Prédiction par apprentissage automatique des évènements critiques présentés par les patients admis aux urgences : de la prédiction pré-hospitalière à l'optimisation du parcours de soin

Sonia Rafi

COLLEGE SCIENCES
DOCTORAL DE LA VIE
BRETAGNE ET DE LA SANTE

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITE DE RENNES

ÉCOLE DOCTORALE N° 637
Sciences de la Vie et de la Santé
Spécialité : *Analyse et Traitement de l'information et des Images Médicales*

Par
Sonia RAFI

Prédiction par apprentissage automatique des évènements critiques présentés par les patients admis aux urgences.

De la prédiction préhospitalière à l'optimisation du parcours de soin.

Thèse présentée et soutenue à Rennes, le 29/11/2023
Unité de recherche : UMR Inserm 1099
Thèse N° :

Rapporteurs avant soutenance :

Leslie Guillon-Grammatico	PUPH Université de Tours
Erik-André Sauleau	PUPH Université de Strasbourg

Composition du Jury :

Président :	Leslie Guillon-Grammatico	PUPH Université de Tours
Examinateurs :	Erik-André Sauleau	PUPH Université de Strasbourg
	Frédéric Holweck	MCU HDR UTBM-Institut Carnot Bourgogne
	Aurélie Bannay	MCUPH Université de Lorraine
	Sahar Bayat- Makoei	Enseignante-Chercheure EHESP Rennes

Dir. de thèse :	Marc Cuggia	PUPH Université de Rennes
Co-dir. de thèse :	Guillaume Bouzillé	MCUPH Université de Rennes
Co-encadrant de thèse :	Cédric Gangloff	MD-PhD Médecine d'Urgence Université de Rennes

Sommaire

1. Introduction générale

1.1. Contexte

L'amélioration de la qualité des soins constitue un enjeu important pour les professionnels de santé. Une prise en charge efficiente des patients admis aux urgences repose sur la capacité à identifier rapidement ceux qui présentent un risque élevé d'évènements critiques. Cette catégorisation dénommée « triage » constitue une composante fondamentale de la médecine d'urgence et impacte directement la morbimortalité des patients (1).

Les équipes médicales doivent évaluer le risque d'aggravation des patients afin de leur proposer un niveau de surveillance approprié à l'issue de leur passage aux urgences : retour à domicile, suivi dans un service standard de médecine ou de chirurgie, admission en unité de surveillance continue ou en réanimation. Cette tâche mobilise une part importante des ressources cognitives des médecins aux urgences et leurs capacités d'analyse sont régulièrement saturées par des flux de données non hiérarchisées et en constante augmentation. À titre d'exemple, un médecin supervise la prise en charge d'environ 30 patients par jour pour lesquels environ 1500 variables sont collectées (informations textuelles, cliniques, constantes, résultats de laboratoire, imagerie), soit environ 45 000 variables par jour. Le dépassement des capacités humaines à traiter ces informations liées à l'afflux de patients, au sous-effectif du personnel soignant et à la fatigue, entraine une augmentation du risque d'erreur. Ce risque pourrait être limité par l'implémentation de systèmes d'aide à la décision basés sur des méthodes utilisant des modèles prédictifs.

Les méthodes de classification supervisée sont susceptibles de répondre à cette problématique en prédisant l'appartenance d'un individu à une classe donnée. Elles permettraient de prédire l'administration de soins critiques aux patients durant leur hospitalisation, incitant à renforcer les moyens humains alloués à la surveillance des patients les plus à risque de présenter un évènement d'intérêt.

L'essor de l'apprentissage automatique ou « machine learning » est lié à deux phénomènes simultanés : l'accroissement de la quantité de données collectées par l'homme et l'augmentation des capacités de calcul en informatique. Cet accroissement a donné lieu à des avancées importantes dans divers domaines : politique, sécurité, sport, agro-alimentaire et audiovisuel. Dans le domaine de la santé, l'imagerie médicale, naturellement orientée vers les technologies du numérique, a joué le rôle de précurseur et constitue toujours l'un des terrains d'application les plus prolifiques pour le développement des systèmes d'aide à la décision basés sur des modèles prédictifs.

Les modèles prédictifs ont fait la preuve de leur efficacité pour résoudre certaines tâches complexes, comme la détection précoce de maladies à partir d'analyses de données d'imagerie. A titre d'exemple, le projet « ChexNet » proposait un modèle d'apprentissage automatique pour le diagnostic de pathologies pulmonaires à partir de radiographies thoraciques (2). Les résultats ont montré que les performances de ce modèle étaient supérieures à celles des médecins radiologues pour la détection de pathologies courantes comme la pneumonie. Ce travail de recherche illustre l'intérêt de l'apprentissage automatique pour la

mise au point de systèmes d'aide à la décision en imagerie médicale et ouvre la voie à une généralisation de ces technologies dans le champ de la santé. Ainsi, il est légitime de penser qu'aucun obstacle théorique n'est susceptible d'entraver la transposition de ces méthodes au domaine particulier des soins critiques.

Cependant, l'intégration de l'apprentissage automatique dans la prise de décision médicale soulève des questions éthiques, de responsabilité et de protection des données. Une implémentation adaptée aux valeurs humanistes inhérentes à chaque système de soins est nécessaire pour s'assurer que ces technologies permettent de manière effective une amélioration de la qualité des soins. En préservant le rôle de l'expertise clinique, les modèles prédictifs sont susceptibles d'impacter de manière positive les performances cliniques et la qualité des soins. Cette réflexion a fait l'objet d'un travail en 2015 par Deo et al. examinant l'impact des avancées technologiques telles que l'augmentation de la puissance de traitement, de la mémoire, du stockage, et l'abondance de données sur l'utilisation de l'apprentissage automatique en médecine (3). Mettant en lumière les succès de l'informatique dans la maîtrise de tâches complexes, les auteurs y ont exploré les possibilités offertes par l'analyse de données massives dans le domaine de la santé, relevant un impact modéré des modèles prédictifs sur les pratiques cliniques en dépit de l'existence de bases de données massives en santé permettant le développement d'algorithmes d'apprentissage automatique.

L'identification précoce des patients nécessitant des soins critiques est un enjeu important pour l'amélioration de la qualité des soins et la pérennisation du système de santé. En effet, le niveau de surveillance proposé à un patient dont l'état clinique est susceptible de se dégrader rapidement correspond à une hospitalisation en réanimation. Ce type de surveillance nécessite une mobilisation importante et continue de moyens humains et techniques disponibles en quantités limitées, et ne peut être proposé de manière systématique à tous les patients hospitalisés. L'intelligence artificielle, en particulier l'apprentissage automatique, ouvre de nouvelles perspectives pour appréhender cette problématique.

Dans ce travail de recherche, nous nous sommes intéressés aux évènements critiques que sont susceptibles de présenter les patients en situation d'urgence. Ces événements d'intérêt ont pour point commun d'engager le pronostic vital des patients. Leur anticipation représente un enjeu clinique, puisque leur délai de détection et de prise en charge constitue un facteur pronostique déterminant en termes de morbimortalité.

On sait par exemple que pour l'arrêt cardiaque le pronostic neurologique est conditionné en premier lieu par le délai de massage cardiaque (4). La survie sans séquelles après arrêt cardiaque en asystolie est de 5 % si le délai d'administration d'adrénaline est inférieur à 2 minutes, ce chiffre baisse d'environ 20 % toutes les 2 minutes, on approche les 100 % de mortalité pour un délai d'administration supérieur à 20 minutes (5).
Dans le choc septique, si l'antibiotique est administré dans l'heure qui suit le début d'hypotension, la mortalité est de moins de 5%, elle est proche de 100% si celui-ci est administré au-delà de 36h du début des signes (6).

Les évènements critiques peuvent également correspondre aux situations d'issues fatales qu'il va être important de prévoir pour le patient et ses proches : décès, soins palliatifs, arrêt des soins.

Concernant les méthodes actuelles de prédiction de ces évènements, il existe de nombreux scores, notamment en réanimation, qui correspondent à des scores de gravité permettant de donner une probabilité de mortalité en fonction du tableau initial.
Par exemple, le score APACHE II constitue un système de classification de la gravité de la maladie appliqué dans les 24 h suivant l'admission d'un patient dans une unité de soins intensifs, ce score qui varie entre 0 et 71 est calculé sur la base de plusieurs mesures : les scores les plus élevés correspondent à un état physiologique décompensé et à un risque de décès élevé (7).
Il fait partie d'une famille de scores de gravité dont les variables constitutives étaient initialement sélectionnées sur choix d'experts. Ces scores ont été améliorés pour aboutir à une deuxième génération dont la sélection et la pondération des variables sont réalisées selon une méthode statistique de régression logistique multivariée et non plus sur dire d'experts (8,9). Les scores de troisième génération intègrent de nouvelles variables et une analyse des anciennes variables selon une prise en charge actualisée des patients de réanimation (10,11). Leur performance a été validée après étude de la discrimination et de la calibration par test d'adéquation du modèle (12,13). Ces scores ont la possibilité d'être ajustés à une sous-population de patients selon une méthode de customisation, en donnant un poids différent à chacune des variables constitutives ou en modifiant globalement la relation entre le score et le risque de décès (14). Finalement, ces scores vont permettre de calculer à l'échelle d'un service une probabilité de mortalité hospitalière en utilisant les données des premières 24h suivant l'admission. En pratique, ils apportent des informations utiles pour évaluer la performance d'un service de manière globale ou dans l'objectif d'obtenir des ressources humaines supplémentaires pour le service concerné du fait de la présence d'une proportion importante de patients graves. En revanche, ces scores ne sont pas suffisamment sensibles pour être utilisés à l'échelle individuelle d'un patient pour une décision de triage, d'admission en réanimation, ou d'abstention thérapeutique et l'utilisation de ces scores comme aide à la décision est consensuellement non autorisée (15,16). Des analyses complémentaires, prospectives, multicentriques, et méthodologiquement plus fiables, restent nécessaires si l'on souhaite appliquer ce type de scores de manière généralisée aux patients de réanimation, qui plus est, aux patients admis aux urgences (17).
D'autres scores ont été développés pour les patients hospitalisés comme le score MEWS qui est un score d'alerte précoce de dégradation clinique des patients basé sur 5 paramètres vitaux : fréquence cardiaque, fréquence respiratoire, pression artérielle systolique, température et réponse neurologique. Il a été créé en 2016 et illustre le début du développement d'outils d'alertes cliniques avec l'objectif de valider prospectivement ce type de méthodes et d'en étudier l'applicabilité clinique (18).

En définitive, nous sommes face à une problématique de prédiction d'évènements graves. Les méthodes utilisées pour y répondre jusqu'à présent en soins critiques et en médecine d'urgence sont globalement des méthodes de régression. Les méthodes d'apprentissage automatique correspondent à un continuum puisqu'elles ont pour objectif, tout comme une régression logistique multivariée, de classer correctement les évènements. Mais celles-ci pourraient permettre d'optimiser la précision de prédiction, comme cela a déjà été montré dans plusieurs domaines, c'est-à-dire qu'elles vont produire un pourcentage de « bien classés » plus important qu'une régression logistique classique (19). En effet, avec ces méthodes, un nombre plus important de variables va pouvoir être traité, ainsi que des données plus complexes comme des signaux ou des images (20).

Pour une exploration plus approfondie de l'état de l'art sur la prédiction des évènements critiques chez les patients hospitalisés aux urgences à travers l'apprentissage automatique, la prochaine partie examinera les avancées récentes dans ce domaine. Cette analyse permettra de contextualiser davantage les implications de l'intégration de l'intelligence artificielle et d'introduire nos travaux de recherche sur les évènements critiques d'intérêt que nous nous proposons d'étudier dans cette thèse.

1.2. État de l'art

Dès le début des années 2000, les premières versions de la base de données Medical Information Mart for Intensive Care (MIMIC) ont vu le jour. MIMIC est une base de données anonymisée qui a été créée en 2001 et a évolué au fil du temps pour devenir un outil de recherche de plus en plus optimisé en médecine intensive. MIMIC I en était la première version (21,22). Elle était constituée d'une collection d'enregistrements multiparamétriques de 90 patients admis en unité de soins intensifs (USI) au Beth Israel Deaconess Medical Center (BIDMC) à Boston entre 1992 et 1999. Les versions suivantes MIMIC II, III, et IV constituent des bases de données dont la taille est plusieurs fois supérieure à la base de données d'origine MIMIC I. D'autres caractéristiques en plus de la taille ont permis leur enrichissement progressif.
MIMIC II, publiée en 2010, comprenait les données de plus de 30000 patients admis dans l'USI du BIDMC entre 2001 et 2008 mais l'accès à cette base de données était réservé à certains professionnels de la recherche dans le domaine sur acceptation d'un dossier de candidature (23). Les données étaient collectées à partir du système Carevue®.
MIMIC III, publiée en 2016, comprend des données de santé anonymisées de plus de 40000 patients admis dans l'USI du BIDMC entre 2001 et 2012 (24). Cette version est publique, accessible gratuitement à l'ensemble des chercheurs. Elle englobe une population diversifiée et très large de patients en soins intensifs et contient des données multiparamétriques caractérisées par leur haut niveau de granularité et la haute résolution temporelle notamment pour les données d'enregistrements au chevet du patient. Les données sont collectées à partir des systèmes Carevue® et Metavision®, qui sont des systèmes d'information qui archivent et affichent les données au chevet des patients.

De 2001 à 2019, le regroupement des sources de données issues des systèmes d'information au chevet des patients avec celles issues des dossiers médicaux (paramètres démographiques, décès en réanimation, à l'hôpital, en dehors de l'hôpital, mesures physiologiques horodatées en réanimation, observations cliniques initiales et d'évolution, prescriptions de traitements, résultats biologiques, comptes-rendus d'imagerie, codages de maladies, de diagnostics, de procédures …) , a permis d'aboutir à la dernière version MIMIC IV qui contient les données de plus de 50000 patients (25).
MIMIC IV a été publiée en 2021. Elle collecte les données des patients admis entre 2008 et 2019 à partir du système Metavision®. Elle représente une évolution majeure par rapport aux versions I, II, III.
En effet, elle comporte des données concernant l'ensemble des séjours hospitaliers des patients du BIDMC sur la période, et non pas seulement les séjours en USI. De plus, elle est séparée en cinq modules afin d'identifier systématiquement la provenance des données :

- Le module « hosp » contient les données hospitalières.
- Le module « ICU » contient les données au niveau des soins intensifs, ce sont les tables d'évènements et leur structure est identique à MIMIC III.
- Le module « ed » contient les données du service d'urgence.
- Le module « cxr » contient les tables de recherches et les métadonnées de MIMIC CXR (fichier d'images et rapports de radiologie).
- Le module « note » contient les notes cliniques anonymisées en texte libre.

Finalement, cette base de données conçue pour être accessible au public vise à soutenir une variété importante d'études et offre du matériel pédagogique contribuant à la réalisation de travaux de recherche dans le domaine d'intérêt.

Plusieurs équipes ont d'ores été déjà travaillé sur la prédiction d'évènements critiques dont nous pouvons détailler quelques exemples avec différentes populations d'étude.

Dès 2012, l'étude d'Escobar et al. visait à développer un modèle prédictif permettant de détecter les transferts non planifiés des patients depuis l'unité médico-chirurgicale vers l'unité de soins intensifs ou le décès sur l'unité en utilisant les données provenant du dossier médical électronique complet de chaque patient. Les résultats ont montré que ce modèle avait une capacité de prédiction élevée avec un coefficient de concordance à 0,84 dans le groupe de dérivation et de 0,78 dans le groupe de validation, ce qui a démontré la faisabilité de la détection précoce de la détérioration des patients en dehors de l'unité de soins intensifs (26).

Plusieurs études ont examiné la prédiction par apprentissage automatique de la septicémie, situation qui peut être à l'origine d'une défaillance hémodynamique grave et rapidement mortelle. En 2018, Nemati et al. ont développé un modèle de prédiction de la septicémie en unité de soins intensifs (27). Ce modèle présentait une précision élevée et permettait de mieux comprendre les facteurs prédictifs de la septicémie dont la correction permettrait d'améliorer le pronostic des patients. C'est également le cas de Desautels et al. qui ont utilisé des données de dossiers médicaux informatiques pour prédire la septicémie en unité de soins intensifs : malgré une approche minimaliste avec un nombre limité de variables, leur modèle a montré une précision élevée et une capacité à prédire

la septicémie plusieurs heures avant l'apparition de signes cliniques (28). Cette même équipe a ensuite réalisé une avancée significative dans une étude dont l'objectif était de prédire les réadmissions non planifiées en unité de soins intensifs, problématique cruciale en termes de prévention des complications et de réduction de la morbimortalité (29). Ce qui distingue particulièrement cette étude, c'est l'utilisation d'une approche de transfert de connaissances : en plus des données de l'hôpital cible au Royaume-Uni, les chercheurs ont également intégré des données de la base de données MIMIC-III, élargissant ainsi leur cohorte et permettant de développer un modèle plus robuste (24). Ce modèle permettait de distinguer les patients présentant un risque élevé de réadmission ou de décès dans les 48 h suivant leur sortie de l'unité. Les résultats de cette étude ont été prometteurs, avec une aire sous la courbe ROC moyenne de 0,7095. Cette performance s'est avérée supérieure à celle du score SWIFT spécialement conçu pour cette tâche, démontrant ainsi la capacité de l'apprentissage automatique à apporter une valeur ajoutée dans la prédiction des réadmissions non planifiées en soins intensifs (30). Enfin, l'étude menée par Calvert et al. en 2019 a exploré par apprentissage automatique la septicémie selon le prisme de la réponse immunitaire à une infection. Cette étude multicentrique a utilisé une vaste base de données regroupant près de 500 000 dossiers médicaux. L'objectif était de développer un outil de diagnostic de la septicémie pour un groupe de patients à haut risque avec un ensemble minimal de variables cliniques. Les modèles obtenus surpassaient les systèmes de notation de gravité classiques ainsi que les biomarqueurs de la septicémie (31).

Il existe également une vaste littérature sur la thématique des arrêts cardiaques. Les modèles d'apprentissage automatique ont été développés pour identifier des marqueurs précoces de cet événement dont le pronostic est intimement lié à la rapidité de réalisation des manœuvres de réanimation cardiopulmonaire. En 2016, Churpek et al. se sont concentrés sur la prédiction des arrêts cardiaques survenant dans les services conventionnels d'hospitalisation (32). Cette recherche basée sur une base de données multicentrique, compare plusieurs techniques d'apprentissage automatique à la régression logistique classique. Les résultats montrent que les méthodes d'apprentissage automatique (en particulier le modèle de forêt aléatoire), surpassent la régression logistique en termes de précision dans la prédiction d'évènements critiques tels que l'arrêt cardiaque, le transfert en unité de soins intensifs, ou le décès. Layeghian et al. ont plus particulièrement étudié cette prédiction en 2018 sur la population spécifique de patients en sepsis (33). Ils retrouvaient une meilleure performance que les scores usuels utilisés en médecine intensive tels les scores APPACHE (11) et MEWS (18).

Pour rester dans le domaine de la cardiologie sans aller jusqu'à l'arrêt cardiaque, un modèle basé sur des réseaux de neurones convolutifs a été développé par Li D. et al. en 2019 pour la reconnaissance des arythmies cardiaques à partir de signaux d'électrocardiogrammes (34). Cette approche a montré une précision de détection élevée avec une moyenne de 99,03 % pour la classification multiclasse, 99,50 % pour la reconnaissance des extrasystoles ventriculaires (ESV), et 99,59 % pour la reconnaissance des extrasystoles auriculaire (ESA). C'est-à-dire qu'avec la méthodologie développée dans cet article, le modèle de réseaux de neurones était capable de classer correctement dans plus de 99% des cas les

signaux ECG comme étant ou non de vraies arythmies, de vraies ESV, de vraies ESA, par rapport à l'ensemble des signaux ECG. Cette méthode pourrait avoir des implications importantes pour la surveillance en temps réel et le diagnostic des maladies cardiaques en particulier lorsqu'elle est combinée avec des technologies portables permettant ainsi d'étendre la détection des arythmies au-delà de l'environnement hospitalier traditionnel. Pour finir une étude récente menée par l'équipe de Gismondi en 2021 a exploré l'utilisation d'algorithmes d'apprentissage automatique pour analyser les images de perfusion myocardique, une technique cruciale dans la détection des maladies coronariennes (35). Les résultats de cette étude ont montré que les modèles d'apprentissage automatique pouvaient distinguer de manière remarquable les cartes de perfusion normale de celles montrant des anomalies. Plus précisément ces modèles ont atteint une précision de plus de 90 % dans cette tâche avec les forêts aléatoires.

Ces différentes études ont utilisé diverses méthodologies, notamment des modèles d'apprentissage automatique supervisés, des analyses de cas témoins, des réseaux de neurones, et parfois des approches avec utilisation minimaliste des données. Elles touchent à plusieurs spécialités et les résultats ont montré que ces modèles pouvaient améliorer la prise de décision clinique dans des domaines critiques tels que la septicémie, les arrêts cardiaques et de manière générale la détérioration physiologique des patients en hospitalisation. Ces avancées dans le domaine de l'apprentissage automatique et de l'intelligence artificielle ont le potentiel de transformer la manière dont les soins de santé sont administrés aux patients en améliorant la réactivité et l'efficacité des interventions des professionnels de soins. Cependant, il est important de prendre en compte les défis liés à l'interprétabilité des modèles et à l'intégration de ces technologies dans les pratiques cliniques.

1.3. Problématique, objectif et déroulement de la recherche

L'objet de ce travail de thèse était donc la prédiction par apprentissage automatique des évènements critiques que peuvent présenter les patients admis aux urgences, avec pour perspective la mise au point d'aide à la décision médicale permettant de détecter les patients à potentiel élevé d'aggravation.

À l'issue du travail d'état de l'art sur le sujet, nous nous sommes retrouvés face à différents verrous :

- des verrous méthodologiques concernant la population d'étude pouvant être soit trop spécifique, soit trop hétérogène, mais aussi concernant les méthodes elles-mêmes avec le problème de l'effet boîte noire en apprentissage automatique et avec les réseaux de neurones profonds. En effet, ces méthodes peuvent aboutir à de bonnes performances, mais ne garantissent pas l'applicabilité clinique.
- des verrous techniques face à l'hétérogénéité et la qualité des données et en premier lieu l'accessibilité des données.
- des verrous humains dans l'application même du fruit du projet

En particulier, concernant la question de l'usage en clinique de ces méthodes par les cliniciens, la réflexion mérite d'être développée à ce stade.
D'abord, il faut prendre en compte la fatigue engendrée par l'excès d'alertes ou alarmes pour les soignants. Plusieurs publications décrivent cette problématique qui correspond à une diminution de l'attention des soignants dans les suites de déclenchements intempestifs de fausses alarmes avec, à terme et de façon paradoxale, un effet délétère sur la prise en charge des patients (36,37). Ceci justifiait d'avoir pour objectif de construire des modèles les plus spécifiques possibles. Dans la même ligne, il est indispensable de créer des modèles robustes avec un cadre d'utilisation clair et précis pour les cliniciens afin d'assurer la confiance dans l'utilisation de ces outils (38).
Par ailleurs, l'utilisation de méthodes d'apprentissage automatique implique la collecte et le partage d'un grand nombre de données, ceci pose la question du respect de la vie privée et de la confidentialité des données des patients (39). Et dans l'optique de l'utilisation même de ces outils par le clinicien, leur complexité et leur manque de transparence peut s'avérer incompatible avec la pratique médicale, dans le cadre de l'information et du consentement éclairé du patient. Notons que quand nous avons débuté ce travail en 2019, l'intelligence artificielle en santé n'avait pas encore fait son entrée dans le code de santé publique, et ça n'est qu'à partir d'août 2021 qu'un cadre juridique a introduit une obligation d'information à la charge des professionnels utilisant une intelligence artificielle en santé (40).

Finalement, la question était celle de pouvoir construire un outil d'aide à la décision performant en clinique pour prédire la survenue d'évènements critiques tels que l'arrêt cardiorespiratoire, la défaillance hémodynamique, le sepsis sévère, le recours aux catécholamines, la détresse respiratoire aiguë, l'intubation en urgence, la perte de vigilance, ou tout simplement l'admission en réanimation, la décision de limiter les thérapeutiques ou le décès. Nous avons fait l'hypothèse que nous pouvions répondre par l'affirmative à cette question, en tirant partie de toutes les données de la trajectoire du patient, ce qui constitue l'originalité de notre approche.
Le projet s'est déroulé en quatre grandes étapes méthodologiques : l'intégration des données, l'extraction de l'information, la construction des modèles, l'évaluation de ces modèles. Et les différents verrous identifiés au départ ont progressivement été levés. L'entrepôt de données eHOP a été utilisé pour la construction des modèles (41).

Cette thèse est composée de trois parties principales, chacune correspondant à un article abordant une problématique spécifique d'événement critique à prédire pour les patients admis aux urgences.
L'article 1, intitulé « Out-of-hospital cardiac arrest detection by machine learning based on the phonetic characteristics of the caller's voice » présente une étude originale sur la prédiction de l'arrêt cardiaque extra hospitalier à l'aide de techniques de machine learning appliquées sur les caractéristiques phonétiques de l'appelant au centre 15. Nous mettons l'accent sur l'importance d'une intervention précoce de la réanimation cardiopulmonaire pour améliorer le pronostic des patients.
L'article 2, intitulé « Machine learning is the key to diagnose COVID-19 : a proof-of-concept study. » aborde spécifiquement l'application des méthodes de

machine learning pour la détection du COVID-19 parmi les patients se présentant aux urgences. En effet, nous avons eu l'occasion d'expérimenter notre méthodologie dans le contexte particulier de cette pandémie qui a constitué un cas d'étude inhabituel. L'objectif était de développer des modèles prédictifs pour faciliter le tri des patients avec l'enjeu clé d'éviter la contamination d'autres patients et de soignants. Nous présentons les résultats de cette étude et discutons de son potentiel dans la gestion des urgences en période de pandémie. L'article 3, intitulé « Predicting critical care scenarios in hospitalized patients using machine learning models » constituera le point essentiel de cette thèse puisqu'il s'attelle à la prédiction des évènements critiques à risque de décès plus ou moins imminent. L'étude est menée sur un nombre important de patients et offre des résultats plus qu'encourageants en termes de précision et de performance. Nous présentons les résultats de cette étude et expliquons comment celle-ci constituera une première étape pour la prédiction des évènements critiques en temps réel pour les patients admis aux urgences.

Nous abordons donc dans chacun de ces 3 articles une problématique spécifique correspondant à un type d'évènement critique d'intérêt dont il nous apparaît important d'en développer la prédiction pour améliorer la prise en charge des patients. Le point commun est la méthodologie utilisée pour prédire ces évènements basés sur l'apprentissage automatique. Nous développons les points communs et la contribution globale de ces 3 articles à la suite de ceux-ci. En fin de compte notre objectif est de fournir des outils performants d'aide à la décision, capables de prédire les évènements critiques et de guider les professionnels de santé dans leurs décisions cliniques.

2. Partie I : Article 1

Rafi S, Gangloff C, Paulhet E, Grimault O, Soulat L, Bouzillé G, Cuggia M. Out-of-Hospital Cardiac Arrest Detection by Machine Learning Based on the Phonetic Characteristics of the Caller's Voice.
Stud Health Technol Inform. 2022 May 25;294:445-449. doi: 10.3233/SHTI220498. PMID : 35612119.

2.1. Résumé de l'article 1

Introduction : l'arrêt cardiaque extra hospitalier (ACEH) constitue un problème majeur de santé publique. Son pronostic est étroitement lié au délai de retour à une activité cardiaque spontanée. Or, les manœuvres de réanimation cardiopulmonaire sont souvent débutées à la suite d'un appel téléphonique au centre 15 sur la consigne de professionnels formés pour identifier les situations critiques par téléphone. Cependant, 25 % des ACEH ne sont pas reconnus lors du premier appel. Il serait donc intéressant de développer des systèmes informatiques automatisés capables de reconnaître ces ACEH lors d'un appel téléphonique. L'objectif de cette étude était la construction et l'évaluation de modèles d'apprentissage automatique permettant de reconnaître l'ACEH en se basant sur les caractéristiques phonétiques de la voix de l'appelant.
Méthodes : tous les patients pour lesquels un appel a été passé au Service d'Aide Medical Urgente (SAMU) 15 du CHU de Rennes, France, entre le 01/01/2017 et le 01/01/2019 étaient éligibles. La variable prédite était la présence d'ACEH. Les variables prédictives ont été collectées par analyse phonétique automatisée de l'appel. Elles se basaient sur les paramètres vocaux suivants : fréquence fondamentale, formants, intensité, Jitter, Shimmer, rapport harmoniques/bruit, nombre de coupures de voix et nombre de périodes. Trois modèles ont été construits utilisant : régression logistique binaire, forêts aléatoires, réseaux de neurones. L'aire sous la courbe (AUC) était le critère de jugement principal pour évaluer les performances de chaque modèle.
Résultats : 820 patients ont été inclus dans l'étude. Le meilleur modèle pour prédire l'ACEH était la forêt aléatoire. (AUC = 74,9 IC95 %= [67,4 - 82,4])
Conclusion : les modèles d'apprentissage automatique basés sur les caractéristiques acoustiques vocales de l'appelant peuvent aider à la reconnaissance de l'ACEH. L'intégration des paramètres identifiés dans cette étude contribuera à la conception de systèmes d'aide à la décision visant à améliorer la détection de l'ACEH par téléphone.

Mots-clés : acoustique, intelligence artificielle, centre d'appel, arrêt cardiaque, opérateur, apprentissage automatique, phonétique, réanimation, analyse vocale.

2.2. Article 1

Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220498

Out-of-Hospital Cardiac Arrest Detection by Machine Learning Based on the Phonetic Characteristics of the Caller's Voice

Sonia RAFI[a]; Cedric GANGLOFF[a]; Etienne PAULHET[b]; Ollivier GRIMAULT[cd]; Louis SOULAT[b]; Guillaume BOUZILLÉ[a]; Marc CUGGIA[a]
[a] *Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, F-35000 Rennes, France*
[b] *Department of Emergency Medicine, CHU Pontchaillou, Rennes, France*
[c] *Search And Rescue team, Department of Emergency Medicine, CHU de la Cavale Blanche, Brest, France*
[d] *EA4324-ORPHY, Université de Bretagne Occidentale, UFR Sciences et Techniques, Brest, France*

Abstract. Introduction. Out-of-hospital cardiac arrest (OHCA) is a major public health issue. The prognosis is closely related to the time from collapse to return of spontaneous circulation. Resuscitation efforts are frequently initiated at the request of emergency call center professionals who are specifically trained to identify critical conditions over the phone. However, 25% of OHCAs are not recognized during the first call. Therefore, it would be interesting to develop automated computer systems to recognize OHCA on the phone. The aim of this study was to build and evaluate machine learning models for OHCA recognition based on the phonetic characteristics of the caller's voice. **Methods.** All patients for whom a call was done to the emergency call center of Rennes, France, between 01/01/2017 and 01/01/2019 were eligible. The predicted variable was OHCA presence. Predicting variables were collected by computer-automatized phonetic analysis of the call. They were based on the following voice parameters: fundamental frequency, formants, intensity, jitter, shimmer, harmonic to noise ratio, number of voice breaks, and number of periods. Three models were generated using binary logistic regression, random forest, and neural network. The area under the curve (AUC) was the primary outcome used to evaluate each model performance. **Results.** 820 patients were included in the study. The best model to predict OHCA was random forest (AUC=74.9, 95% CI=67.4-82.4). **Conclusion.** Machine learning models based on the acoustic characteristics of the caller's voice can recognize OHCA. The integration of the acoustic parameters identified in this study will help to design decision-making support systems to improve OHCA detection over the phone.

Keywords. cardiac arrest, resuscitation, dispatcher, call center, acoustic, phonetic, voice analysis, artificial intelligence, machine learning.

1. Introduction

Out-of-hospital cardiac arrest (OHCA) is a major public health concern [1]. The prognosis is closely related to the time from collapse to return of spontaneous circulation [2]. The resuscitation efforts are frequently initiated at the request of emergency call center professionals who are specifically trained to identify critical conditions over the phone. However, 25% of OHCAs are not recognized during the first call, most often because

emergency call centers are overwhelmed [3]. In this context, it would be interesting to develop automated computer systems to recognize OHCA based on the bystanders' speech on the phone.

Speech analysis can be decomposed in two fields: linguistic and phonetic. Linguistic analysis investigates the meaning of words and their relationships, while phonetic analysis focuses on the voice acoustic characteristics. Acoustic analysis is based on the following principle: the acoustic signal is generated in the glottis and passes through the vocal tract where it is modulated by the pharyngeal, buccal, labial and nasal cavities acting as filters [4]. The different frequency bandwidths emitted at the end of the vocal tract are called "formants". The human voice is composed of one fundamental frequency (F0) and four formants (F1 to F4) that correspond to each of the four cavities. Formant frequencies vary over time in function of the spatial conformation changes of the cavities driven by the phonatory muscles. Other characteristics, such as intensity variations and amount of noise contained in the acoustic signal, also are taken into account in the phonetic analysis. Software tools for fast and automated phonetic analysis have been recently developed [5].

The aim of this study was to build and evaluate machine learning models that can recognize OHCA based on the phonetic characteristics of the caller's voice.

2. Methods

The study protocol was approved by the Medical Ethics Committee of Rennes academic hospital (approval number 19.116, issued on December 4, 2019).

2.1. Software

Acoustic features of the recorded calls were extracted with WC-MDX Workstation version 11.6.0.0, UHERS Corporation, 2005. Phonetic analyses were performed with PRAAT v6.1.03, 2019, Institute of Phonetic Sciences, Amsterdam University. All statistical analyses and model building were performed with "R-studio", version 1.3.1093, RStudio PBC, 2009-2021. The following R packages were used: "Dplyr", version 1.0.0, for data manipulation; "MICE", version 3.9.0, for missing data implementation; and "Caret", version 6.0-90, for model building.

2.2. Setting and study population

Data were collected retrospectively from patients for whom a call was done to the emergency call center of Rennes academic hospital, France, between 01/01/2017 and 01/01/2019.

2.3. Variables and groups

The predicted variable was OHCA presence. Patients included in the study were divided in two groups: i) OHCA group if they had OHCA, and ii) NO-OHCA group, if they did not have a diagnosis of OHCA.

Predicting variables were collected by automated phonetic analysis. They were based on the following parameters: fundamental frequency, formants, intensity, jitter, shimmer,

harmonic to noise ratio, number of voice breaks, and number of periods. A full description of all these variables is available in the PRAAT software documentation [5]. Briefly, the spectrogram of the human voice includes five high intensity bands. One corresponds to the fundamental frequency and the other four to formants (figure 1). The *fundamental frequency* (F0) is determined by the tension of the vocal cords and the subglottic pressure, and varies with the stress level. *Formant frequencies* (F1 to F4) are determined by the resonance system volume. *Jitter* measures the short-term variations in the fundamental frequency in seconds, and *shimmer* reflects short-term disturbances in signal intensity. The *noise to harmonic ratio* is defined by the ratio of the non-harmonic and harmonic intensities contained in the acoustic signal. The *number of voice breaks* is the number of distances between consecutive pulses that are longer than a defined duration divided by the pitch floor. The *number of periods* is calculated by counting the number of time intervals between glottal pulses.

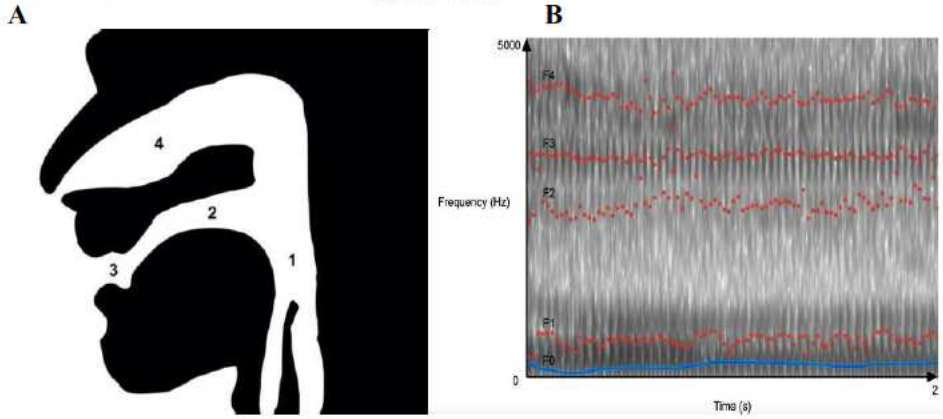

Figure 1: Bases of the automated phonetic analysis. A: Voice signal. The acoustic signal is generated in the glottis and modulated by four resonance systems: the pharyngeal (1), buccal (2), labial (3), and nasal (4) cavities. Each cavity acts as a filter, and the bandwidths emitted at the end of the vocal tract are called "formants". **B: Voice spectrogram.** The human voice is composed of one fundamental frequency (F0, blue line) and four formants (F1 to F4, red dots). Formants vary over time according to the spatial conformation changes of the four cavities driven by the phonatory muscles. A spectrogram is the graphic representation of these five bandwidths in function of time.

2.4. Statistics

The Student's *t*-test was used to compare means between groups. A p-value <0.05 was considered significant. To avoid multicollinearity, correlation coefficients were calculated for each pair of variables, and when the coefficient was higher than 0.8, one variable was excluded. *Data splitting*: Data were randomly divided in two parts: training dataset and test dataset. The training dataset corresponded to 80% of the complete dataset and was used to build the models. *Model training*: Three models were constructed: binary logistic regression, random forest (500 trees), and neural network (3 layers). *Performance measurement*: The predictions made by the three models were compared to the "OHCA" variable in the test dataset and receiver operating characteristic curves (ROC) were constructed accordingly. The area under the curve (AUC) was the primary outcome used to evaluate each model performance.

3. Results

3.1. Selected patients

820 patients were included in the study, 410 in each group.

3.2. Predicting variables

Table 1 shows the comparison of the mean values of each selected predicting variable between groups.

Table 1. Comparison of the selected predicting variables between groups. Values in brackets represent the 95% confidence intervals. *t-test significance was set at p <0.05 level. OHCA= Out of Hospital Cardiac Arrest, med=median, min=minimum, max=maximum, sd=standard deviation, n=number, NHR= noise to harmonic ratio.

Variable	OHCA group (n=410)	NO-OHCA group (n=410)	p (t-test)
Pitch mean (Hz)	244 (238 - 250)	197 (192 - 202)	< 0.001*
Pitch sd (Hz)	45.3 (43.5 - 47)	47 (44.9 - 49.1)	0.216
Pitch min (Hz)	136 (131 - 140)	109 (106 - 113)	< 0.001*
Pitch max (Hz)	417 (409 - 425)	395 (386 - 404)	< 0.001*
N of voice breaks	12.2 (11.9 - 12.5)	12.5 (12.2 - 12.9)	0.181
Jitter local absolute	$8.5e^{-5}$ ($8.2e^{-5}$ - $8.9e^{-5}$)	$10.7e^{-5}$ ($10.2e^{-5}$ - $11.2e^{-5}$)	< 0.001*
Jitter RAP (%)	0.82 (0.80 - 0.84)	0.84 (0.81 - 0.87)	0.399
Shimmer local (dB)	1.01 (1.00 - 1.03)	1.03 (1.01 - 1.04)	0.335
Shimmer APQ11 (%)	10.0 (9.7 - 10.2)	9.9 (9.6 - 10.3)	0.930
Mean NHR	0.153 (0.147 - 0.158)	0.154 (0.147 - 0.161)	0.741
Formant med H1 (Hz)	518 (512 - 525)	493 (487 - 498)	< 0.001*
Formant med H2 (Hz)	1482 (1469 - 1495)	1454 (1441 - 1468)	< 0.001*
Formant med H3 (Hz)	2288 (2279 - 2298)	2281 (2271 - 2290)	0.247
Formant med H4 (Hz)	3063 (3057 - 3069)	3074 (3068 - 3081)	0.014*
Intensity med (dB)	68.3 (67.6 – 69.0)	65.2 (64.5 - 65.9)	< 0.001*
Intensity sd (dB)	14.0 (12.9 - 15.2)	11.1 (10.8 - 11.4)	< 0.001*
Intensity min (dB)	27.6 (26.6 - 28.6)	27.7 (27.0 - 28.4)	0.887
Intensity max (dB)	79.6 (79.1 - 80.1)	78.2 (77.7 - 78.8)	< 0.001*

3.3. Model performance analysis

Table 2 shows the performance of the three models.

Table 2. ROC-AUC value for OHCA prediction by each model. Models used the acoustic features of the caller's voice to predict OHCA. Values in brackets represent the 95% confidence intervals.

Model	ROC-AUC
Binary logistic regression	71.4 (63.5-79.4)
Random forest	74.9 (67.4-82.4)
Neural network	64.5 (57.3-71.8)

4. Discussion

This study describes machine learning models that use the acoustic features of the bystander's voice to predict OHCA. These acoustic features were previously identified as stress markers. In 2010, Frampton and al. observed significant variations in pitch and

fundamental frequency in employees facing time restrictions to complete an order [6]. Similarly, Mendoza et al. [7] reported an increase in fundamental frequency, jitter and shimmer perturbations in individuals subject to work-related stressful conditions. To our knowledge, our study is the first to show similar results in the field of emergency healthcare. We observed statistically significant differences in the bystanders' voice acoustic parameters in the presence of OHCA (table 1). We also demonstrated that it is possible to create decision-making support models to recognize OHCA based on these parameters (table 2). These results provides perspectives for short-term applications of machine learning models that integrate semantic and acoustic parameters. Indeed, in 2019, Blomberg et al. developed a model based only on semantic elements that could reduce OHCA detection time [8]. The integration of the acoustic parameters described in the present study should increase the performance of such models.

5. Conclusion

This study demonstrates that machine learning models can recognize OCA based on the acoustic characteristics of the caller's voice. The integration of the acoustic parameters identified in this study could help to increase the performance of decision-making support systems that already integrate semantic parameters in order to improve OHCA detection over the phone.

References

[1] Berdowski J, Berg RA, Tijssen JGP, Koster RW. Global incidences of out-of-hospital cardiac arrest and survival rates: Systematic review of 67 prospective studies. Resuscitation 81, 1479–1487 (2010).
[2] Rossetti AO, Oddo M, Logroscino G, Kaplan PW. Prognostication after cardiac arrest and hypothermia: A prospective study. Ann. Neurol. 67, 301–307 (2010).
[3] Viereck S, Møller TP, Rothman JP, Folke F, Lippert FK. Recognition of out-of-hospital cardiac arrest during emergency calls — a systematic review of observational studies. Scand. J. Trauma Resusc. Emerg. Med. 25, 9 (2017).
[4] Titze IR. Some Consensus has been Reached on the Labeling of Harmonics, Formants, and Resonances. J. Voice 30, 129 (2016).
[5] Praat: doing Phonetics by Computer. https://www.fon.hum.uva.nl/praat/.
[6] Frampton M, Sripada S, Bion RAH, Peters S. Detection of time-pressure induced stress in speech via acoustic indicators.
[7] Mendoza E. Carballo G. Acoustic analysis of induced vocal stress by means of cognitive workload tasks. J. Voice Off. J. Voice Found. 12, 263–273 (1998).
[8] Blomberg SN, et al. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation 138, 322–329 (2019).

2.3. Résultats obtenus et discussion

Dans ce premier article, nous nous sommes donc penchés sur un événement critique de toute première importance : l'arrêt cardiaque extra hospitalier (ACEH). L'idée de cet article est venue de plusieurs constats.
Tout d'abord, l'ACEH constitue un enjeu majeur de santé publique par sa fréquence et par la gravité de son pronostic. Chaque année 0,5 à 1 pour 1000 habitants dans le monde serait victime d'arrêt cardiaque inopiné (42). Malgré les progrès réalisés dans la prise en charge de l'ACEH, la survie reste faible, inférieure à 10 % (43). Parmi ces survivants, l'évolution neurologique est variable, allant de la survie sans séquelles avec fonctions cognitives normales, à l'état végétatif. En effet, le résultat neurologique est intimement lié au délai de retour à une activité cardiaque spontanée (RACS), qui lui-même va découler du délai d'initiation des manœuvres cardiopulmonaires (44).
Ces manœuvres de réanimation s'inscrivent dans le processus de chaîne de la survie, largement développée dans les politiques de santé publique ces dernières décennies (45). Elles doivent être idéalement débutées par le premier témoin de l'événement sur les lieux, qui souvent va être la personne qui va contacter les secours. Les centres d'appels d'urgence (centre 15 en France) jouent un rôle crucial en identifiant les situations critiques sur la base des éléments fournis par l'appelant au téléphone.
Dans cette situation, nous avons constaté que malgré l'utilisation d'algorithmes décisionnels validés dans la reconnaissance de ces ACEH, une proportion non négligeable, environ 30 %, n'étaient pas reconnus au premier appel (46). Ainsi, il fallait réfléchir à des méthodes pour améliorer cette reconnaissance, et proposer de nouveaux systèmes d'aide au diagnostic pour les professionnels de santé exerçant dans ces centres.
Nous avons pensé qu'il serait intéressant de développer ces systèmes informatiques automatisés pour reconnaître l'ACEH sur les caractéristiques de l'appel au secours du témoin au téléphone, plus précisément sur la base des caractéristiques phonétiques de sa voix.
Nous avons fait l'hypothèse que les caractéristiques phonétiques de la voix de l'appelant à l'aide seraient modifiées par rapport à la normale, du fait du stress engendré par le fait d'être témoin d'un ACEH et qu'il serait licite de développer des modèles d'apprentissage automatique utilisant comme variables prédictives les paramètres phonétiques de la voix du témoin lors de son premier appel au centre 15.
Pour mener cette étude, il a d'abord fallu poser quelques bases sur l'analyse du discours qui peut se décomposer en deux versants : linguistique liée au vocabulaire, à la sémantique et aux relations entre les mots, et phonétique qui se concentre sur les caractéristiques acoustiques de la voix. L'utilisation des paramètres linguistiques du discours comme variables prédictives pour aider au diagnostic d'arrêt cardiaque par des méthodes d'apprentissage automatique a déjà été étudiée en 2019 par Blomberg et al. avec de bonnes performances des modèles proposés (47).
En pratique, nous avons utilisé l'outil Praat qui est un logiciel Open source de phonétique permettant d'analyser toute donnée sonore. Nous avons ainsi recueilli de façon automatisée les paramètres acoustiques suivants : la fréquence fondamentale, les formants, l'intensité, le Jitter, le Shimmer,

le rapport harmonique/bruit, le nombre de coupures de voix et le nombre de périodes. Les valeurs de ces paramètres ont été analysées sur les 30 premières secondes de la bande sonore correspondant au discours de l'appelant.
Les paramètres : fréquence fondamentale, Jitter et Shimmer avaient déjà montré des perturbations en dehors du domaine de la santé, dans des situations de stress liées aux conditions de travail (48). À notre connaissance, il n'existait pas d'étude cherchant à montrer des résultats similaires dans le domaine des soins en médecine d'urgence, dans la situation où un témoin appelle pour signaler un problème potentiellement grave notamment un ACEH.

Nous avons donc construit trois modèles d'apprentissage automatique : régression logistique binaire, forêt aléatoire et réseau neuronal. Ils correspondent aux modèles habituellement utilisés pour répondre aux questions de prédiction en médecine d'urgence. L'aire sous la courbe (AUC) Receiver Operating Characteristics (ROC) a été utilisée pour évaluer la performance de chaque modèle.
Le modèle de régression logistique a été réalisé avec la fonction train du package caret (method = "glm" et family = binomial). Cela indique que nous faisons une régression logistique binomiale : nous avons utilisé la fonction glm (generalized linear model) avec l'argument family=binomial pour effectuer une régression logistique. Cette fonction « standard » ne fait pas appel à des techniques comme le lasso ou l'elastic net.
La forêt aléatoire (Random forest) est également implémentée avec la fonction train du package caret. Ici, nous avons utilisé method = « rf », ce qui signifie que le modèle est basé sur l'approche de forêt aléatoire.
Le troisième modèle est un réseau de neurones artificiels, spécifié par method = « pcaNNet ». Cela implique un réseau de neurones avec analyse en composantes principales pour la réduction de dimensionalité.
Chacun des modèles utilisait le même ensemble de variables prédictives et était entraîné sur le même jeu de données.
Le jeu de données initial comportait 34 variables dont 18 ont été sélectionnées pour la modélisation qui figurent dans le tableau 1 de l'article.

Les résultats de cette étude ont montré que la plupart des variables acoustiques étudiées présentaient une différence significative entre le groupe « ACEH » et le groupe « non ACEH », avec une pertinence clinique sur la variable fréquence fondamentale qui correspond à un marqueur de stress dans le spectrogramme vocal. En termes de performances, le meilleur modèle pour prédire l'ACEH était la forêt aléatoire avec une AUC proche de 0,75.
Ceci suggérait que les caractéristiques phonétiques de la voix de l'appelant pouvaient fournir des indices prédictifs significatifs de l'ACEH. Cette conclusion revêt une importance particulière dans cette situation où chaque minute compte dans l'introduction de la réanimation cardiopulmonaire pour restaurer une circulation efficace et améliorer le pronostic neurologique du patient. Notre étude a apporté des éclairages importants pour l'amélioration de la détection précoce de l'ACEH via les

appels au centre 15 et a eu l'originalité de se baser non pas sur les paramètres cliniques du patient mais sur les caractéristiques de la voix de l'appelant qui sont mises en relation avec le niveau de stress et donc avec la probabilité d'être en situation d'arrêt cardiaque pour le patient. Cette étude ouvre des champs d'applications à court terme, et des modèles d'apprentissage automatique intégrant à la fois des paramètres sémantiques et acoustiques pourraient rapidement être développés. En effet, le modèle publié en 2019 par Blomberg basé uniquement sur des éléments sémantiques a montré une réduction du temps de détection des ACEH (47). L'intégration des paramètres acoustiques décrits dans la présente étude devrait augmenter les performances de tels modèles.

En mettant ces résultats en perspective avec le cadre plus général du sujet de la thèse centré sur la prédiction des évènements critiques aux urgences, nous pouvons faire plusieurs observations.
Tout d'abord, l'utilisation de l'apprentissage automatique pour analyser les caractéristiques phonétiques de la voix de l'appelant au centre 15 constitue une approche novatrice et prometteuse. Cette approche s'ajoute à l'arsenal croissant d'outils basés sur l'intelligence artificielle pour améliorer la détection et la prise de décision clinique dans les urgences.
Ensuite, cette étude souligne l'importance de l'intégration de données multiples dans les modèles de prédiction. En utilisant des informations non-médicales, telles que les caractéristiques phonétiques de la voix, les chercheurs ont réussi à améliorer la précision de la détection de l'ACEH. Cela rappelle l'importance de la prise en compte d'une variété de paramètres dans les modèles de prédiction clinique, ce qui pourrait s'avérer essentiel pour obtenir des performances optimales.
Enfin, l'étude a mis en évidence le potentiel des systèmes d'alerte automatisés basés sur l'apprentissage automatique pour améliorer la reconnaissance des situations critiques en médecine d'urgence. L'automatisation de la détection de l'ACEH pourrait à terme permettre de libérer des ressources humaines dans les centres d'appels d'urgence, permettant aux professionnels de santé de se concentrer sur des tâches plus spécialisées. Cela pourrait avoir un impact considérable sur la qualité des soins aux patients critiques, et améliorer de façon significative la morbimortalité.

Toutefois, il est important de noter certaines limites de cette étude.
En premier lieu, elle a été menée sur un seul centre et l'échantillon des patients inclus était relativement restreint, ce qui pourrait limiter la généralisation des résultats à d'autres populations. L'événement ACEH reste en soi un événement rare, comparativement à l'ensemble des évènements médicaux pouvant survenir dans une population donnée, ce qui donne lieu à un effectif restreint : ceci se traduit par une performance correcte, mais qui mérite d'être améliorée. De plus, les résultats de cette étude ne sont pas suffisamment précis et ne comptaient pas suffisamment d'observations pour permettre d'identifier un profil de voix comme devant particulièrement alerter du risque d'ACEH.
Il faudrait également probablement prendre en compte plus précisément le profil de l'appelant en considérant qui est cette personne par rapport à

la victime : un adulte, un enfant, un parent, un aidant, un professionnel de soin, ou un simple témoin anonyme, car probablement que ces paramètres exercent une influence sur la valeur des paramètres vocaux analysés lors de ces appels. Des études futures sur des échantillons plus vastes et plus variés seront nécessaires pour valider la robustesse des modèles de prédiction proposés.
Ensuite, bien que les caractéristiques phonétiques de la voix de l'appelant se soient révélé être des indicateurs potentiels d'ACEH, les modèles que nous avons développés dans cette étude ne peuvent être, en l'état, substitués aux évaluations médicales traditionnelles. L'ACEH est un événement médical complexe avec de nombreux facteurs de risque et de nombreux paramètres prédictifs potentiels. Les caractéristiques phonétiques de la voix peuvent être utiles pour améliorer la détection précoce, mais elles ne peuvent pas remplacer les compétences cliniques des professionnels de santé, car un nombre important d'ACEH n'a pas été identifié dans cette étude. Néanmoins, de tels algorithmes pourraient être utiles comme outils d'aide à la décision pour réduire l'erreur humaine en alertant les professionnels de santé lorsqu'une situation potentiellement grave est détectée.
Enfin, l'intégration des méthodes d'apprentissage automatique dans l'exercice des professionnels de santé en centre d'appel d'urgence nécessiterait des efforts supplémentaires pour garantir la confidentialité et la sécurité des données du dossier médical des patients. Une validation et une adaptation plus spécifique à chaque centre d'appel seraient indispensables.

En conclusion, cette première étude a présenté une contribution significative à la prédiction des évènements critiques aux urgences en introduisant une approche innovante basée sur les caractéristiques phonétiques de la voix de l'appelant au centre d'appel d'urgence. Les résultats montrent que ces caractéristiques peuvent être utilisées pour améliorer la détection précoce de l'ACEH. La méthodologie doit encore être améliorée pour obtenir de meilleures performances si l'on souhaite l'utiliser en pratique clinique. Mais cette approche s'inscrit dans le contexte plus large de l'utilisation de l'apprentissage automatique pour améliorer la prise de décision clinique et l'efficacité des systèmes d'alerte précoce aux urgences. Les perspectives d'application de ces méthodes pourraient avoir un impact considérable sur la gestion des urgences médicales et l'amélioration des résultats des patients présentant des évènements critiques lors de leur hospitalisation après l'admission aux urgences.

3. Partie II: Article 2

Gangloff C, Rafi S, Bouzillé G, Soulat L, Cuggia M. Machine learning is the key to diagnose COVID-19: a proof-of-concept study.
Sci Rep. 2021 Mar 30 ;11(1) :7166. doi : 10.1038/s41598-021-86735-9.

3.1. Résumé de l'article 2

Le test d'amplification en chaîne par polymérase à transcription inverse (RT-PCR) est la norme acceptée pour le diagnostic de la maladie du coronavirus 2019 (COVID-19). Comme tout test, la RT-PCR peut fournir des résultats faussement négatifs qui peuvent être rectifiés par les cliniciens en confrontant les données cliniques, biologiques et d'imagerie. Le fait d'associer la RT-PCR et la tomodensitométrie (TDM) thoracique pourrait améliorer les performances diagnostiques, mais cela nécessiterait des ressources considérables pour une utilisation rapide chez tous les patients suspects de COVID-19. La contribution potentielle de l'apprentissage automatique dans cette situation n'a pas encore été évaluée. L'objectif de cette étude était de développer et d'évaluer des modèles d'apprentissage automatique utilisant des données cliniques et biologiques de routine pour améliorer les performances de la RT-PCR et de la TDM thoracique dans le diagnostic du COVID-19 chez les patients hospitalisés après une prise en charge aux urgences. Tous les adultes admis aux urgences pour suspicion de COVID-19, puis hospitalisés au centre hospitalier universitaire (CHU) de Rennes, France, entre le 20 mars 2020 et le 05 mai 2020, ont été inclus dans l'étude. Trois types de modèles ont été créés : régression logistique, forêt aléatoire et réseau neuronal. Chaque modèle a été entrainé à prédire le diagnostic de COVID-19 à partir de différentes variables prédictives.
Après généralisation, les modèles d'apprentissage automatique permettront d'améliorer les performances de la TDM thoracique et de la RT-PCR dans le diagnostic du COVID-19.

3.2. Article 2

scientific reports

OPEN

Machine learning is the key to diagnose COVID-19: a proof-of-concept study

Cedric Gangloff[1]✉, Sonia Rafi[1], Guillaume Bouzillé[1], Louis Soulat[2] & Marc Cuggia[1]

The reverse transcription-polymerase chain reaction (RT-PCR) assay is the accepted standard for coronavirus disease 2019 (COVID-19) diagnosis. As any test, RT-PCR provides false negative results that can be rectified by clinicians by confronting clinical, biological and imaging data. The combination of RT-PCR and chest-CT could improve diagnosis performance, but this would requires considerable resources for its rapid use in all patients with suspected COVID-19. The potential contribution of machine learning in this situation has not been fully evaluated. The objective of this study was to develop and evaluate machine learning models using routine clinical and laboratory data to improve the performance of RT-PCR and chest-CT for COVID-19 diagnosis among post-emergency hospitalized patients. All adults admitted to the ED for suspected COVID-19, and then hospitalized at Rennes academic hospital, France, between March 20, 2020 and May 5, 2020 were included in the study. Three model types were created: logistic regression, random forest, and neural network. Each model was trained to diagnose COVID-19 using different sets of variables. Area under the receiving operator characteristics curve (AUC) was the primary outcome to evaluate model's performances. 536 patients were included in the study: 106 in the COVID group, 430 in the NOT-COVID group. The AUC values of chest-CT and RT-PCR increased from 0.778 to 0.892 and from 0.852 to 0.930, respectively, with the contribution of machine learning. After generalization, machine learning models will allow increasing chest-CT and RT-PCR performances for COVID-19 diagnosis.

The severe acute respiratory syndrome coronavirus 2 (SARS-coV-2) outbreak started in December 2019 in the Hubei province, China. The associated disease, coronavirus disease 2019 (COVID-19)[1], has now spread worldwide. The World Health Organization currently reports more than 10 million confirmed cases and 500,000 deaths. Increased mortality rates and the collapse of healthcare systems have been reported in several regions[2–4]. Indeed, due to SARS-coV-2 contagiousness, promiscuity within health systems can promote patient-to-patient transmission[5,6] and the contamination of healthcare workers[7], rapidly leading to the saturation of health systems[8]. To limit this effect, patients with COVID-19 infection are hospitalized in specific units after being emergency department (ED) triage[9]. Therefore, it is essential to have a reliable and easy-to-use tool for COVID-19 diagnosis. SARS-coV-2 real-time RT-PCR reverse transcription–polymerase chain reaction (RT-PCR) is the accepted standard for COVID-19 diagnosis[10]. However, RT-PCR performances are sub-optimal and, like for any other test, there are false negatives results[11,12]. Therefore, additional investigations should be performed in patients with negative RT-PCR results but high clinical probability of COVID-19. In this context, chest-CT is an interesting tool because it allows detecting virus-induced lung tissue damages and alternative diagnoses[13]. Thus, when a patient presents a high clinical probability of COVID-19, a negative RT-PCR and a chest-CT showing typical COVID-19 lesions with no sign of alternative diagnosis, it is possible to consider that the patient has COVID-19 with a false negative RT-PCR result. The use of chest-CT alone cannot be recommended, but its combined use with clinic and RT-PCR allows to resolve diagnostic ambiguities[14]. However, RT-PCR and chest-CT cannot be performed in all patients suspected to have COVID-19 for many reasons, including reagent shortage[15], device unavailability, lack of human resources, and high costs. Moreover, the time required to perform both tests increase the risk of ED overcrowding by patients waiting for their results. Therefore, health professionals must adapt their diagnostic strategies in function of their resources[16]. To our knowledge, the potential contribution of machine learning using imaging, clinical and laboratory data has been poorly evaluated in this context. Machine learning is an inherited artificial intelligence approach that enables computers to extract or classify patterns. It allows predicting whether a patient belongs to a predefined group using explanatory variables. The recent increase in machine learning models in the healthcare field suggests that these methods could improve the COVID-19

[1]Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, F-35000 Rennes, France. [2]Department of Emergency Medicine, CHU Rennes, F-35000 Rennes, France. ✉email: cedric.ganglof@gmail.com

diagnostic strategy[17]. The objective of this study was to develop and evaluate machine learning models using clinico-biological data from health records to improve the RT-PCR and chest-CT performances for COVID-19 diagnosis among post-emergency hospitalized patients.

Materials and methods

This study protocol was approved by the Medical Ethics Committee of Rennes academic hospital (approval number 0020.93 issued on July 7, 2020). All methods were performed in accordance with the relevant guidelines and regulations. Authorization to conduct research from the Clinical Data Warehouse of Hospital of Rennes was given by CNIL—Commission Nationale Informatique et Liberté (Authorization number 2020-028 issued on February 27, 2020). Informed written consent from each participant was not required for this study according to the French Data Protection Act of 6 January 1978, as this study only included information from existing medical records and did not involve interaction with patients or collection of identifiable private information. Each entry of sample data was deidentified to ensure confidentiality.

Software. Data extractions, manipulations, statistical analyses, and model buildings were performed with "R-studio", version 1.3.1093, RStudio PBC, 2009–2020. Specialized packages and functions were used for specific analysis: "Dplyr", version 1.0.0 was used for data manipulation, "Purrr", version 0.3.4 for data simplification, and "missForest", version 1.4 for missing data imputation. Variable importance calculations and K-fold cross-validation were performed with the "Caret" package, version 6.0-86. Correlations matrix were calculated with the "corrplot" package, version 0.84. Random forests were built with "randomForest" version 4.6-14 and artificial neural networks with "neuralnet" version 1.44.2. "pROC" version 1.16.2 was used to generate the receiver operating characteristic (ROC) curves and calculate the area under the curve (AUC) for each model.

Setting. Data were collected retrospectively from patients admitted to the adult E.D. of Rennes Academic Hospital, France.

Patient selection. All post-emergency hospitalized patients ≥ 18 years old admitted between March 20, 2020 and May 5, 2020 and suspected to have COVID-19 were included in the study.

Data collection. Data were automatically collected from "eHOP", a local clinical data warehouse in which health data are integrated and de-identified in real time[18]. Structured data, such as laboratory results, were directly collected from the data warehouse. Text fields were structured by using regular expressions[19].

Data pre-processing. In the raw data-frame, all values were associated with a unique identifier (ID) corresponding to each patient's admission. This data-frame contained multiple lines per ID (Fig. 1, step 1). Variables collected more than once during the patient journey appeared as lists (Fig. 1, step 2). Lists were simplified according to the type of variable (Fig. 1, step 3).

Predicted variable. The predicted variable for each patient was the presence of COVID-19. COVID-19 was diagnosed as follows. Patients were triaged at ED admission and considered "suspected COVID-19" when they had at least one symptom compatible with COVID-19. Symptoms considered as compatible with COVID-19 were the followings: cough, dyspnea, hyperthermia, myalgias, asthenia, diarrhea, confusion or anosmia. After triage, patients were examined by an ED physician who estimated the clinical probability of COVID-19 (low, intermediate or high) and the need for hospitalization. RT-PCR and chest-CT were performed in all hospitalized patient suspected to have COVID-19. When suspected COVID-19 patients had positive RT-PCR, they were considered "COVID-19 positive", regardless of the level of clinical probability. When clinical probability was "high" and chest-CT showed typical COVID-19 images with no sign of alternate diagnosis, the patient was considered "COVID-19 positive", even if RT-PCR was negative. In this case, the RT-PCR result was considered a false negative[20,21]. Patients were allocated to "COVID" and "NOT-COVID" groups accordingly.

Predicting variables selection. All clinical and laboratory variables present in the database were collected. The Student's *t*- and chi-square tests were used to compare means between groups for numerical and binary variables, respectively. A p value < 0.05 was considered statistically significant. Variables with a p value < 0.2 were considered variables of interest. To avoid multicollinearity, correlation coefficients were calculated for each pair of variables of interest. When correlation coefficient was higher than 0.8, one of the two variables was excluded.

Data split. Data were randomly divided in two parts: the train data-frame, and the test data- frame. The train data-frame corresponded to 80% of the whole data-frame and was used to build the models. Models performances were evaluated using the test data-frame that corresponded to the remaining 20%.

Missing data imputation. Before the training process, missing values were imputed independently for each data-frame with a non-parametric procedure developed by Stekhoven and Buhlmann. This method called "missForest" is well-suited for mixed datasets requiring categorical and continuous variables imputations and is based on a random forest model trained iteratively. In this method, an evaluation is made after each iteration by calculation of the normalized root mean squared error and implementation is stopped when the evaluation indicates a decrease in performance. Three iteration were performed with 100 tree per random forest in this study.

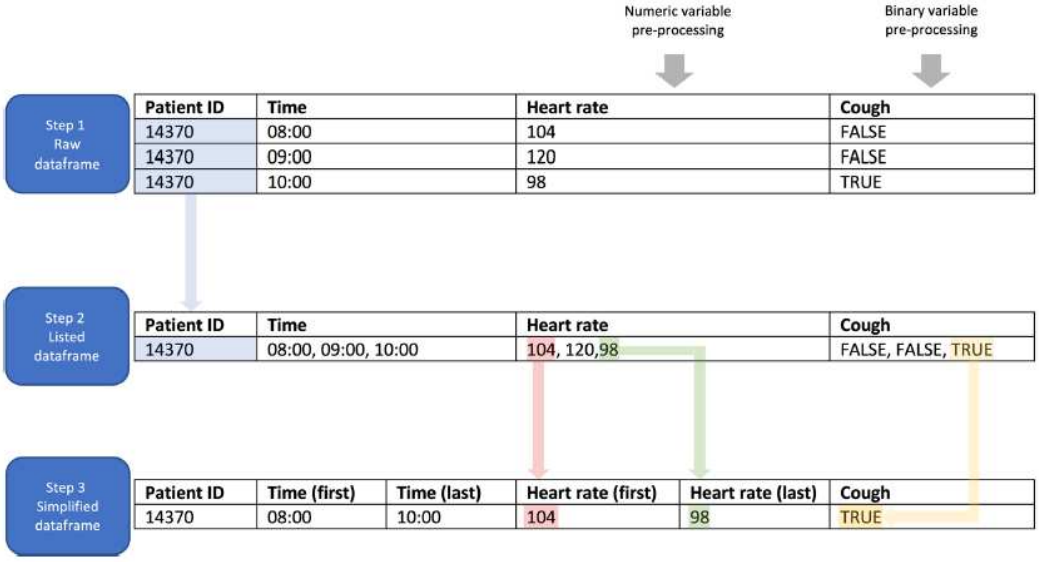

Figure 1. Data pre-processing. *The first step* corresponded to raw data, as they were initially stored in the database. Each ID was characterized by multiple rows. *On the second step*, data were listed in chronological order, with a single row per ID (blue arrow). *In the third step*, data were simplified. For numeric variables, only the first value was selected (red arrow). For binary variables, the value "true" was retained when it was present at least once in the list (yellow arrow).

Model training. Three model types were constructed: binary logistic regressions, random forests and artificial neural networks. Random forest models were trained with 500 trees; neural networks were composed of three layers. Each model type was trained with three sets of variables: clinico-biological variables, clinico-biological variables with chest-CT, and clinico-biological variables with RT-PCR. A k-fold cross validation was performed in order to prevent over-fitting. Overfitting occurs when à machine learning algorithm captures the noise of the data. In this case, high performances are observed on the training data, but poor results are observed on new data. In other words, overfitted models cannot give suitable predictions on new patients. K-folds cross validation is a high-performance method to prevent overfitting. In this approach, the data-frame is divided into k parts called "folds". A model is trained by using k – 1 folds, and the remaining fold is used to validate the model. The same procedure is applied k times (once per fold). This approach is well-suited for small datasets, but requires more calculations. In this study, k = 10 folds were used to build the models.

Variable importance. To compare the importance of the different variables, the value of the most important variable in each model was arbitrarily set at 100 and the relative importance of each variable was determined with an adequate method depending on the model. In binary logistic regressions, the absolute value of the t-statistic for each parameter was used to calculate the importance of each variable. In random forests, the prediction error on the out-of-bag portion of the data was recorded for each tree and he same was done after permuting each predictor variable. The difference between the two were averaged over all trees and normalized by the standard deviation of the differences to determine each variables importance. In the neural network models, the method was based on combinations of the absolute values of the weights[22].

Performance measurement. Models were built with the train data-frame and their performances were assessed on the test data-frame, whose data were not used for model-building. This procedure guarantees non-biased performances measurements by confronting the models to unseen data, as if they were challenged to predict the presence of COVID-19 among new patients. The area under ROC curves is commonly used to evaluate and compare classifiers in machine learning, biomedical and bioinformatics applications[23]. In this study, models' predictions were compared to the "COVID" variable in the test data-frame and ROC curves were constructed accordingly. The AUC was the primary outcome used to evaluate each model performance.

Ethics approval. This study was approved by the ethic committee of Rennes academic hospital (number of approval: 20.93).

Consent for publication. All methods were performed in accordance with the relevant guidelines and regulations. Authorization to conduct research from the Clinical Data Warehouse of Hospital of Rennes was

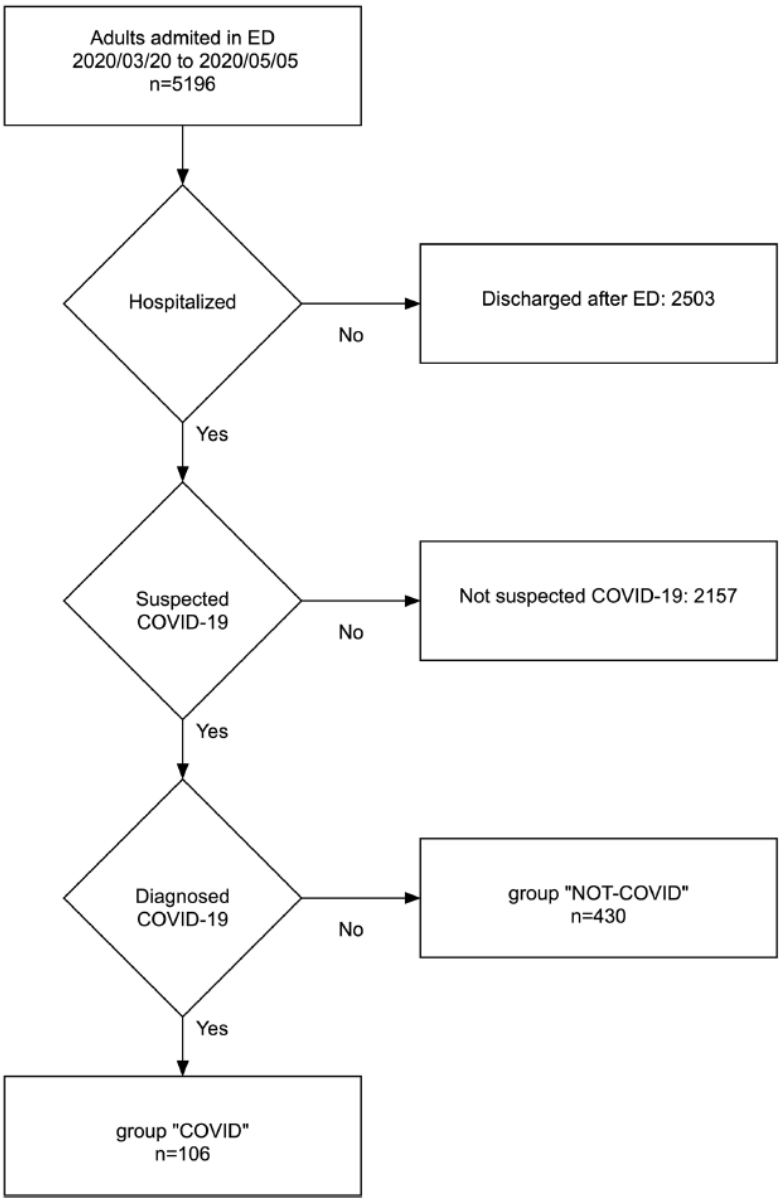

Figure 2. Flow chart of patient selection. Patients suspected to have COVID-19 had at least one of the following symptoms: cough, dyspnea, hyperthermia, myalgias, asthenia, diarrhea, confusion or anosmia. Both chest-CT and RT-PCR were performed in all patients with suspected COVID-19 who were hospitalized.

given by CNIL—Commission Nationale Informatique et Liberté (Authorization number 2020-028 issued on February 27, 2020). Informed written consent from each participant was not required for this study according to the French Data Protection Act of 6 January 1978, as this study only included information from existing medical records and did not involve interaction with patients or collection of identifiable private information. Each entry of sample data was deidentified to ensure confidentiality.

Diagnostic	n (%)
COVID-19	106 (19.8)
Cardiac insufficiency	98 (18.3)
Pneumonia	74 (13.8)
Chronic obstructive pulmonary disease (COPD)	52 (9.7)
Influenza-like illness	38 (7.1)
Intra-abdominal infection	34 (6.3)
Asthma	20 (3.7)
Non organic dyspnea	19 (3.5)
Urinary tract infection	19 (3.5)
Confusion in the elderly (delirium)	14 (2.6)
Transcient fever	14 (2.6)
Cancer	12 (2.2)
Pulmonary embolism	12 (2.2)
Skin infection	6 (1.1)
Others	5 (0.9)
Central nervous system infection	4 (0.7)
Heart infection (pericarditis, myocarditis, endocarditis)	4 (0.7)
Prothesis-related infection	3 (0.6)
Traumatic dyspnea	2 (0.4)
Total	536 (100)

Table 1. Diagnostic categories for the 536 suspected COVID-19 hospitalized patients. All patients presented at last one clinical sign compatible with COVID-19 and underwent chest-CT and RT-PCR. 106 were classified in the COVID group, 430 in the NOT-COVID group.

Results

Patient selection. The patient selection flow chart is presented Fig. 2.

Diagnostics. Diagnostics for the 536 patients selected in this study are represented in Table 1.

Selected variables. Twenty-three clinico-biological variables were considered as variables of interest (Table 2). Variables not selected as variables of interests are presented in supplementary Table 1.

Variables correlations. Calculation of the correlation coefficients for each pair of the 23 variables of interest (Fig. 3) showed that two variables were highly correlated with a correlation coefficient > 0.8: neutrophil count and leukocyte count. Leukocyte count was removed from model building. Therefore, the final set of clinical and laboratory variables selected for model building included 22 variables.

Chest-CT and RT-PCR performances. AUCs of chest-CT and RT-PCR used alone for COVID-19 diagnosis were 0.778 (CI 95% 0.682–0.873) and 0.852 (CI 95% 0.764–0.940), respectively.

Models performance. The AUC values for the three model types trained with each set of variables are presented in Table 3.

The ROC curves for the binary logistic regression models are presented Fig. 4.

Importance of clinico-biological variables. The importance of the different variables in each model type is presented Table 4.

Discussion

Models presented in this study were trained on typical suspected COVID-19 patients. All models were trained and evaluated using data from patients with diseases (e.g. heart failure, pneumonia, asthma, COPD; Table 1) that are frequently observed in ED and that share clinical symptoms with COVID-19. The finding that our machine learning models could differentiate between these diseases and COVID-19 suggests that they could be implemented in other EDs with similar patient populations.

The variables selected for model-building were consistent with the clinico-biological signs of COVID-19. These variables belong to five categories: clinical signs, arterial blood gas, blood cell count, ionogram, hemostasis and liver enzymes. Clinical signs: the proportion of cough, hyperthermia, myalgia, asthenia, diarrhea, and confusion was significantly higher in the COVID-19 than in the NO-COVID-19 group. Such

	NOT-COVID (n = 430)	COVID (n = 106)	P value
Clinicals and treatments			
Cough, %	83.0 (79.1–87.0	92.4 (86.5–98.2)	0.0563
Hyperthermia, %	66.7 (61.8–71.7)	77.2 (67.9–86.4)	0.0940
Myalgias, %	17.1 (13.2–21.1)	34.1 (23.7–44.6)	0.0012*
Asthenia, %	30.9 (26.1–35.8)	45.5 (34.5–56.5)	0.0187*
Diarrhea, %	22.9 (18.5–27.3)	32.9 (22.5–43.2)	0.0867
Confusion, %	21.7 (17.4–26.1)	7.5 (1.7–13.4)	0.0063*
Furosemid (usual treatment), %	16.0 (12.2–19.9)	6.3 (0.9–11.7)	0.0401*
Arterial blood gas			
Base excess, mmol/L	3.0 (2.6–3.4)	2.7 (1.8–3.6)	0.0151*
Lactates, mmol/L	1.7 (1.5–1.9)	1.3 (1.1–1.5)	< 0.001*
Complete blood count			
Red blood cell count, Tera/L	4.2 (4.1–4.3)	4.5 (4.3–4.7)	< 0.001*
Mean platelet volume, fL	8.6 (8.4–8.8)	8.8 (8.5–9.1)	0.0269*
Leukocytes, G/L	10.2 (9.6–10.8)	7.7 (6.7–8.7)	0.0568
Neutrophils, G/L	7.9 (7.4–8.4)	6 (5.1–6.9)	0.1488
Platelet count	236.1 (225.8–246.4)	198.9 (182.1–215.7)	0.0482*
Eosinophils percentage	1.4 (1.1–1.7)	0.8 (0.4–1.2)	0.0873
Basophils percentage	0.6 (0.5–0.7)	0.4 (0.3–0.5)	< 0.001*
Lymphocytes, G/L	1.3 (1.2–1.4)	1 (0.8–1.2)	< 0.001*
Monocytes, G/L	0.8 (0.7–0.9)	0.6 (0.5–0.7)	< 0.001*
Ionogram			
Potassium, mmol/L	4.1 (4–4.2)	4 (3.8–4.2)	0.0039*
Phosphor, mmol/L	1 (0.9–1.1)	1.1 (0.9–1.3)	< 0.001*
Hemostasis and liver enzymes			
Alanine aminotransferase, mmol/L	64.5 (47.1–81.9)	46.2 (33.9–58.5)	0.1845
International normalized ratio	1.3 (1.2–1.4)	1.2 (1.1–1.3)	< 0.001*
D-Dimer, ng/ml	2200 (1600–2800)	2800 (1400–4200)	< 0.001*

Table 2. Variables of interest. Means and percentage between groups were compared with Student's t- and chi-square tests, respectively. Only variables with $p < 0.02$ were considered as variables of interest and were listed in this table. Values in brackets represent the 95% confidence intervals. *$p < 0.005$.

symptoms have previously been reported in numerous studies[24–28]. Interestingly, anosmia was not selected as a variable of interest, suggesting a lack of relevance of this symptom in our setting[29,30]. Arterial blood gas: in the NOT-COVID group, serum lactate concentration was higher, and base-excess was lower than in the COVID group, revealing the presence of patients with circulatory failure, a frequently reported complication of bacteremia[31]. Therefore, serum lactate concentration and base-excess are relevant for differentiating between patients with COVID-19 and with bacterial infections. Blood cell count: the mean leukocyte, lymphocyte, and platelet counts were lower in the COVID than in the NOT-COVID group. Previous authors have reported similar results. Indeed, a meta-analysis from Zhu et al. showed that patients with COVID-19 do not have hyperleukocytosis, except when associated with bacteremia[32]. COVID-19-associated lymphopenia correlates with the disease severity and is related to an immune response deficiency[33]. Similarly, thrombocytopenia was previously identified as a poor prognosis factor in this context[34]. Indeed, a meta- analysis by Lippi et al. revealed that platelet count was significantly lower in patients with severe COVID-19[35], suggesting an inappropriate activation of the coagulation process. Ionogram: the mean potassium concentration was lower in the COVID group. This could be due to hyperventilation, but further investigation must be conducted to confirm this hypothesis. Hemostasis and liver enzymes: the mean D-dimer concentration was higher in the COVID group than in the NOT-COVID group. Elevated D-dimers are associated with higher rates of thromboembolic events[36]. These results are in line with the theory of an increased thromboembolic risk in patients with COVID-19[37–39]. This finding could be associated with the presence of antiphospholipid antibodies, but the pathophysiology of this phenomenon is still debated[40]. Variables selected for model building were therefore consistent with previous studies that have reported clinico-biological signs of COVID-19.

Machine learning models will help to triage COVID-19 patients. RT-PCR and chest-CT are expensive, require qualified professionals to perform them and it is a real challenge to be able to get efficiently these two examinations in the context of a pandemic. An increasing number of patients awaiting results of these tests can lead to ED overcrowding and increased mortality rates in an epidemic context[41,42]. The logistic regression model presented in this study and trained only with clinico-biological variables had an AUC value of 0.754. This model only requires clinical examination and routine biology assays: complete blood cell count, ionogram,

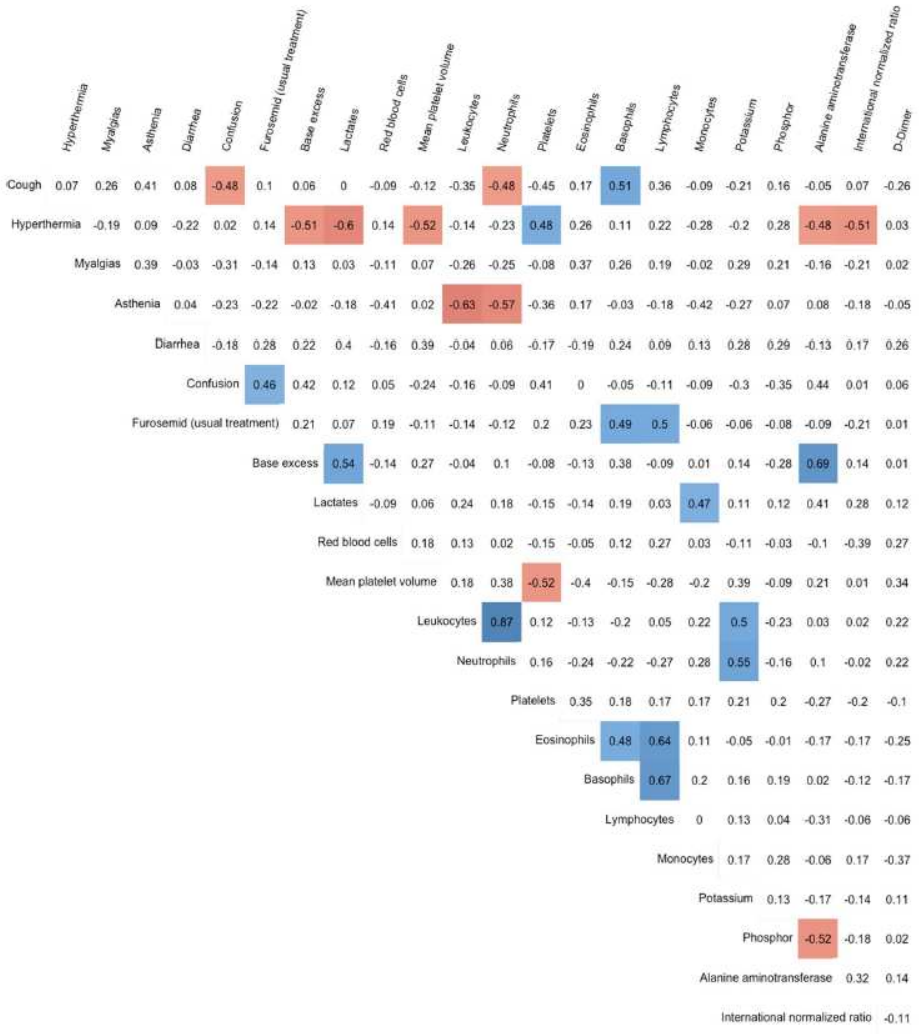

Figure 3. Correlation coefficients for all pairs of variables of interest. Pearson and Spearman coefficients were calculated for continuous and binary variables, respectively. Correlations not significantly different from 0 are in white cells. Positive correlations are in blue cells, and negative correlations in red cells. Leukocyte count and neutrophil count were identified as highly correlated, and leukocyte count was removed from model building.

	Clinico-biological	Clinico-biological + chest-CT	Clinico-biological + RT-PCR
Binary logistic regression	0.772 (0.668–0.875)	0.886 (0.804–0.968)	0.930 (0.867–0.992)
Random forest	0.754 (0.638–0.871)	0.829 (0.724–0.935)	0.903 (0.816–0.989)
Artificial neural network	0.728 (0.617–0.840)	0.892 (0.811–0.973)	0.844 (0.731–0.957)

Table 3. AUC for each machine learning model. Three model types were constructed: binary logistic regression, random forest, and artificial neural network. Each model was trained with three sets of variables: clinico-biological, clinico-biological with chest-CT, and clinico-biological with RT-PCR. Models were built and assessed on two separate data-frames. Values in brackets represent the 95% confidence intervals.

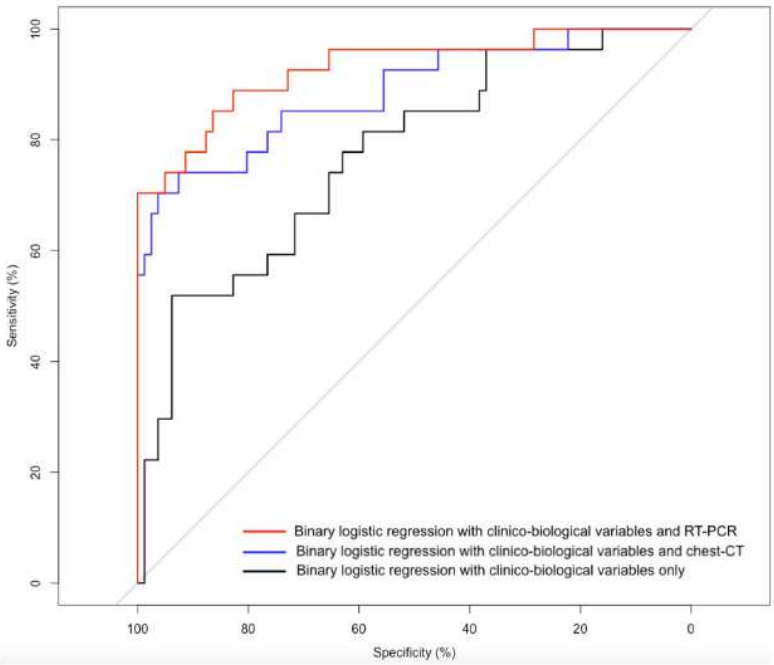

Figure 4. ROC curves for the 3 logistic regression models based on common clinico-biological variables alone, clinico-biological variables with chest-CT and common clinico-biological variables with RT-PCR. The "Binary logistic regression with clinico-biological variables and RT-PCR" was the best performing model in this study.

standard hemostasis tests, liver enzymes, and arterial blood gas. Such tests are low- cost and can be realized worldwide using automated devices. Therefore, in ED, a first triage identifying patients requiring isolation might be done by using machine learning while waiting for the RT-PCR result.

Machine learning will improve RT-PCR and chest-CT performance for COVID-19 diagnosis. Several studies found sub-optimal performances of these tests when only one is used for COVID-19 diagnosis[11,13,43,44]. Indeed, the sensitivity of the RT-PCR test depends on the number of cycles used to determine the cut-off value for positivity[45] and one of the issues by using chest-CT alone for COVID-19 diagnosis is the risk of false positive[20,46]. Artificial intelligence methods could be used to overcome these drawbacks. Some studies have already investigated the use of artificial intelligence for COVID-19 diagnosis and their number is progressively increasing with the pandemic duration[47,48]. Many of these studies are based on deep neural networks to improve COVID-19 diagnosis by chest-CT or X-ray imaging, particularly to help to differentiate between COVID-19 lesions and bacterial lung diseases[49–52]. For examples, the COVID-net tool based on 16,756 chest radiography images across 13,645 patients has an accuracy of 92.4%, the COVID-19 detection neural network (COVNet) based on 4356 chest-CT from 3322 patients has an accuracy of 95%[53,54]. However, few studies used laboratory, clinical, and imaging data together for COVID-19 diagnosis. Our results are in line with studies that used machine learning models based on clinico-biological variables for COVID-19 diagnosis[55–57]. The performances of these models were low, excepted for the model described by Plante and al. that used data from a large sample, but did not include imaging data[58]. Another study integrated RT-PCR, chest-CT and clinico-biological data, like in the present work, but the study population was smaller, and the performance was slightly lower[59]. In our study, the AUC values for chest-CT and RT-PCR increased from 0.778 to 0.892 and from 0.852 to 0.930 with the contribution of machine learning. The generalization of such models will allow increasing the diagnostic performances of both chest-CT and RT-PCR for COVID-19 diagnosis.

Limitations. Our study has some limitations. First, the machine learning models developed in this experimentation are not directly transferrable to other hospitals due to it's monocentric design. Such models must be further developed and tested on a larger scale to be generalized. However, the predictive variables selected and identified as highly important in this study are similar to the clinical and biological signs reported by previous authors, suggesting the absence of major obstacles for model generalization. Second, the study population included only hospitalized patients suspected to have COVID-19. It would be interesting to perform a similar study in non-hospitalized patient to test the model performances for COVID-19 diagnosis in paucisympto-

	Binary logistic regression	Random forest	Artificial neural network
Clinicals and treatments			
Cough	62	1.5	21.3
Hyperthermia	24.7	3.1	24.1
Myalgias	40.5	9.36	39.7
Asthenia	24.1	6.7	34.1
Diarrhea	57.1	6	22.9
Confusion	91.8	5.8	33
Furosemid (usual treatment)	72.8	0	22.3
Arterial blood gas			
Base excess	31.6	37.3	16.2
Lactates	100	100	77.9
Complete blood count			
Red blood cell count	85.1	86.4	41
Mean platelet volume	0	34.5	0
Neutrophils	64.6	50.6	55.1
Platelet count	36.8	46.8	47.3
Eosinophils	35.9	36.5	57.1
Basophils	74.8	79.9	100
Lymphocytes	67.4	51.8	46
Monocytes	27.6	57.3	52.6
Ionogram			
Potassium	17.5	39.4	19.4
Phosphor	6.2	45.5	4.2
Hemostasis and liver enzymes			
Alanine aminotransferase	64.6	39.8	3.1
International normalized ratio	9	39.2	0.2
D-Dimer	57.6	58.4	10

Table 4. Importance of clinico-biological variables by decreasing order in each model type. The relative importance of each variable was calculated in comparison with the most important variable in the model, whose importance was arbitrarily set at 100.

matic patients. Finally, the classification of chest-CT as negative on the basis of the absence of typical images of COVID-19 might need to be reviewed in line with recent publications on COVID-19 diagnosis using deep learning methods.

Conclusion. Our study demonstrates that machine learning models can be developed for improving COVID-19 diagnosis in patients hospitalized through the ED. Models based on chest-CT or RT-PCR will increase the performance of these tests by using clinico-biological variables. After generalization, machine learning should play a key role in the management of the outbreak by improving the performances of chest-CT and RT-PCR for COVID-19 diagnosis.

Data availability

After publication, the data will be made available to others on reasonable requests to the corresponding author.

Received: 30 October 2020; Accepted: 16 March 2021
Published online: 30 March 2021

References

1. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): A review. *JAMA* **324**, 782 (2020).
2. Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of Novel coronavirus-infected pneumonia. *N. Engl. J. Med.* **382**, 1199–1207 (2020).
3. Korean Society of Infectious Diseases and Korea Centers for Disease Control and Prevention. Analysis on 54 mortality cases of coronavirus disease 2019 in the Republic of Korea from January 19 to March 10, 2020. *J. Korean Med. Sci.* **35**, e132 (2020).
4. Peng, L. *et al.* Improved early recognition of coronavirus disease-2019 (COVID-19): Single-center data from a Shanghai Screening Hospital. *Arch. Iran. Med.* **23**, 272–276 (2020).
5. Wong, S. C. Y. *et al.* Risk of nosocomial transmission of coronavirus disease 2019: An experience in a general ward setting in Hong Kong. *J. Hosp. Infect.* **105**, 119–127 (2020).
6. For the Singapore 2019 Novel Coronavirus Outbreak Research Team *et al.* Detection of air and surface contamination by SARS-CoV-2 in hospital rooms of infected patients. *Nat. Commun.* **11**, 2800 (2020).

7. Xiao, J., Fang, M., Chen, Q. & He, B. SARS, MERS and COVID-19 among healthcare workers: A narrative review. *J. Infect. Public Health* **13**, 843–848 (2020).
8. Coccolini, F. *et al.* COVID-19 the showdown for mass casualty preparedness and management: The Cassandra Syndrome. *World J. Emerg. Surg.* **15**, 26 (2020).
9. Maves, R. C. *et al.* Triage of scarce critical care resources in COVID-19 an implementation guide for regional allocation. *Chest* **158**, 212–225 (2020).
10. Hanson, K. E. *et al.* Infectious diseases society of America Guidelines on the Diagnosis of COVID-19. *Clin. Infect. Dis.* https://doi.org/10.1093/cid/ciaa760 (2020).
11. Xiao, A. T. False negative of RT-PCR and prolonged nucleic acid conversion in COVID-19: Rather than recurrence. 2.
12. Li, Y. *et al.* Stability issues of RT-PCR testing of SARS-CoV-2 for hospitalized patients clinically diagnosed with COVID-19. *J. Med. Virol.* **92**, 903–908 (2020).
13. Ai, T. *et al.* Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases. 23.
14. Rubin, G. D. *et al.* The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the Fleischner Society. *Radiology* **296**, 172–180 (2020).
15. Beetz, C. *et al.* Rapid large-scale COVID-19 testing during shortages. *Diagnostics* **10**, 464 (2020).
16. Lone, S. A. & Ahmad, A. COVID-19 pandemic—an African perspective. *Emerg. Microbes Infect.* **9**, 1300–1308 (2020).
17. Furlow, B. Deep learning poised to revolutionise diagnostic imaging. *Lancet Respir. Med.* **5**, 779 (2017).
18. Delamarre, D., Bouzille, G., Dalleau, K., Courtel, D. & Cuggia, M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. 5.
19. Tang, X., Zeng, Q., Cui, T. & Wu, Z. Regular expression-based reference metadata extraction from the web. in *2010 IEEE 2nd Symposium on Web Society* 5607427 (IEEE, 2010). https://doi.org/10.1109/SWS.2010.5607427.
20. Caruso, D. *et al.* Chest CT features of COVID-19 in Rome, Italy. *Radiology* **296**, E79–E85 (2020).
21. Chung, M. *et al.* CT imaging features of 2019 novel coronavirus (2019-nCoV). *Radiology* **295**, 202–207 (2020).
22. Gevrey, M., Dimopoulos, I. & Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.* **160**, 249–264 (2003).
23. Robin, X. *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **12**, 77 (2011).
24. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **395**, 497–506 (2020).
25. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* **395**, 507–513 (2020).
26. Jiang, F. *et al.* Review of the clinical characteristics of coronavirus disease 2019 (COVID-19). *J. Gen. Intern. Med.* **35**, 1545–1549 (2020).
27. Goyal, P. *et al.* Clinical characteristics of covid-19 in New York City. *N. Engl. J. Med.* **382**, 2372–2374 (2020).
28. Wang, D. *et al.* Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061 (2020).
29. Lechien, J. R. *et al.* Olfactory and gustatory dysfunctions as a clinical presentation of mild-to-moderate forms of the coronavirus disease (COVID-19): A multicenter European study. *Eur. Arch. Otorhinolaryngol.* **277**, 2251–2261 (2020).
30. Beltrán-Corbellini, Á. *et al.* Acute-onset smell and taste disorders in the context of COVID-19: A pilot multicentre polymerase chain reaction based case–control study. *Eur. J. Neurol.* **27**, 1738–1741 (2020).
31. Shankar-Hari, M. *et al.* Developing a new definition and assessing new clinical criteria for septic shock: For the third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* **315**, 775 (2016).
32. Zhu, J. *et al.* Clinicopathological characteristics of 8697 patients with COVID-19 in China: A meta-analysis. *Fam. Med. Community Health* **8**, e000406 (2020).
33. Azkur, A. K. *et al.* Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. *Allergy* **75**, 1564–1581 (2020).
34. Hu, L. *et al.* Risk factors associated with clinical outcomes in 323 COVID-19 hospitalized patients in Wuhan, China. 33.
35. Lippi, G., Plebani, M. & Henry, B. M. Thrombocytopenia is associated with severe coronavirus disease 2019 (COVID-19) infections: A meta-analysis. *Clin. Chim. Acta* **506**, 145–148 (2020).
36. Crawford, F. *et al.* D-dimer test for excluding the diagnosis of pulmonary embolism. *Cochrane Database Syst. Rev.* https://doi.org/10.1002/14651858.CD010864.pub2 (2016).
37. Spiezia, L. *et al.* COVID-19-related severe hypercoagulability in patients admitted to intensive care unit for acute respiratory failure. *Thromb. Haemost.* **120**, 998–1000 (2020).
38. Oxley, T. J. *et al.* Large-vessel stroke as a presenting feature of Covid-19 in the young. *N. Engl. J. Med.* **382**, e60 (2020).
39. Zhang, L. *et al.* D-dimer levels on admission to predict in-hospital mortality in patients with Covid-19. *J. Thromb. Haemost.* **18**, 1324–1329 (2020).
40. Zhang, Y. *et al.* Coagulopathy and antiphospholipid antibodies in patients with Covid-19. *N. Engl. J. Med.* **382**, e38 (2020).
41. Geelhoed, G. C. & Klerk, N. H. Emergency department overcrowding, mortality and the 4-hour rule in Western Australia. *Med. J. Aust.* **196**, 122–126 (2012).
42. Kim, J. *et al.* Maximum emergency department overcrowding is correlated with occurrence of unexpected cardiac arrest. *Crit. Care* **24**, 305 (2020).
43. Long, C. *et al.* Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT?. *Eur. J. Radiol.* **126**, 108961 (2020).
44. Liu, R. *et al.* Positive rate of RT-PCR detection of SARS-CoV-2 infection in 4880 cases from one hospital in Wuhan, China, from Jan to Feb 2020. *Clin. Chim. Acta* **505**, 172–175 (2020).
45. Liu, Z. High sensitivity detection of SARS-CoV-2 using multiplex PCR and a multiplex-PCR-based metagenomic method. 24.
46. Yang, W. *et al.* The role of imaging in 2019 novel coronavirus pneumonia (COVID-19). *Eur. Radiol.* **30**, 4874–4882 (2020).
47. Albahri, A. S. *et al.* Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): A systematic review. *J. Med. Syst.* **44**, 122 (2020).
48. Dananjayan, S. & Raj, G. M. Artificial Intelligence during a pandemic: The COVID-19 example. *Int. J. Health Plann. Manag.* **35**, 1260–1262 (2020).
49. Wu, X. *et al.* Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study. *Eur. J. Radiol.* **128**, 109041 (2020).
50. Kang, H. *et al.* Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning. *IEEE Trans. Med. Imaging* **39**, 2606–2614 (2020).
51. Fan, Z., Jamil, M., Sadiq, M. T., Huang, X. & Yu, X. Exploiting multiple optimizers with transfer learning techniques for the identification of COVID-19 patients. *J. Healthc. Eng.* **2020**, 1–13 (2020).
52. Jang, S. B. *et al.* Deep-learning algorithms for the interpretation of chest radiographs to aid in the triage of COVID-19 patients: A multicenter retrospective study. *PLoS One* **15**, e0242759 (2020).
53. Wang, L., Lin, Z. Q. & Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **10**, 19549 (2020).
54. Kumar, A., Gupta, P. K. & Srivastava, A. A review of modern technologies for tackling COVID-19 pandemic. *Diabetes Metab. Syndr. Clin. Res. Rev.* **14**, 569–573 (2020).

55. Goodman-Meza, D. *et al.* A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. *PLoS One* **15**, e0239474 (2020).
56. D'Ambrosia, C., Christensen, H. & Aronoff-Spencer, E. Computing SARS-CoV-2 infection risk from symptoms, imaging, and test data: Diagnostic model development. *J. Med. Internet Res.* **22**, e24478 (2020).
57. Langer, T. *et al.* Development of machine learning models to predict RT-PCR results for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in patients with influenza-like symptoms using only basic clinical data. *Scand. J. Trauma Resusc. Emerg. Med.* **28**, 113 (2020).
58. Plante, T. B. *et al.* Development and external validation of a machine learning tool to rule out COVID-19 among adults in the emergency department using routine blood tests: A large, multicentre, real-world study. *J. Med. Internet Res.* **22**, e24048 (2020).
59. Hermans, J. J. R. *et al.* Chest CT for triage during COVID-19 on the emergency department: Myth or truth?. *Emerg. Radiol.* **27**, 641–651 (2020).

Author contributions

C.G. designed the experiment, collected data, performed statistical analysis and build machine-learning models. C.G. and S.R. wrote the manuscript. G.B., L.S. and M.C. read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-86735-9.

Correspondence and requests for materials should be addressed to C.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

3.3. Résultats obtenus et discussion

Notre article a donc présenté une étude type preuve de concept qui explorait l'utilisation de méthodes d'apprentissage automatique dans le cadre de l'aide au diagnostic du COVID-19 pour les patients hospitalisés après une admission aux urgences. Cette étude a été menée sur les données des patients adultes admis aux urgences du CHU de Rennes durant la « première vague » de la pandémie au printemps 2020, période particulièrement marquée par l'importance de disposer de méthodes diagnostiques rapides et précises afin de lutter contre la propagation de cette maladie pour laquelle les connaissances étaient encore très mal maîtrisées à cette époque.
Rappelons que cette maladie, due au SARS-CoV-2, avait été initialement détectée en décembre 2019 à Wuhan en Chine (49), puis s'est rapidement étendue en Europe, en Amérique du Nord et dans le monde entier avec déjà plus de 200 000 cas recensés par l'OMS à la date du 22/03/2020 (50).
Le taux de létalité était d'emblée non négligeable avec une proportion importante de formes graves (50). Un confinement des populations a par conséquent été décidé dans de nombreux pays pour freiner la progression de cette pandémie, et les activités professionnelles non-indispensables impliquant des regroupements d'individus ont été arrêtées. En revanche, pour des raisons évidentes de santé publique, les hôpitaux ont continué de prendre en charge les patients dont les soins ne pouvaient être reportés (infarctus du myocarde, accident vasculaire cérébral, greffes, dialyse), en accueillant à fortiori également des patients atteints de la COVID-19. Parmi ces patients, ceux qui n'étaient pas diagnostiqués étaient alors susceptibles de contaminer les soignants, entrainant une circulation active du SARS-CoV-2 au sein d'hôpitaux devenant eux-mêmes les foyers de l'épidémie qu'ils étaient censés combattre. Ce mécanisme avait déjà rapidement entrainé en quelques jours l'effondrement du système de santé dans certains territoires, expliquant une augmentation dramatique du taux de mortalité (51). Par conséquent, il était fondamental d'assurer la séparation des patients hospitalisés en deux circuits distincts le plus précocement possible : l'un dédié aux patients atteints de la COVID-19, et l'autre aux patients à priori indemnes de cette maladie. Et pour réaliser ce tri, il était indispensable de disposer d'une méthode diagnostique performante de la COVID-19 avant l'hospitalisation. Le diagnostic reposait en l'état des connaissances sur l'examen clinique, la biologie et l'imagerie. Les symptômes cliniques étaient les suivants : syndrome grippal (toux, asthénie, myalgies), signes digestifs (diarrhées et douleur abdominales) et ORL (agueusie et anosmie). Les signes biologiques étaient la lymphopénie, la thrombopénie, l'élévation de la CRP et une PCR SARS-CoV-2 sur écouvillon nasopharyngé positive. Les images considérées comme évocatrices de COVID-19 étaient des lésions périphériques bilatérales en verre dépoli à prédominance postéro-basale à la TDM thoracique.

À cette date-là, il n'existait pas de test permettant de conclure de manière certaine à la présence ou l'absence de la maladie avant hospitalisation.
C'est-à-dire qu'il n'existait pas de gold standard et deux tests étaient disponibles à l'époque : le 1er était l'imagerie thoracique montrant des plages en verre dépoli bilatérales à prédominance postéro basales ou « crazy paving », le 2ème était la RT-PCR qui s'est imposée depuis pour des raisons

de faisabilité. A cette époque, la performance de ces tests n'était pas connue dans la littérature, nous avons donc choisi de considérer qu'un patient était atteint de COVID-19 lorsque l'un de ces deux tests était positif.

Ainsi, nous avons pensé que les méthodes de classification supervisée développées dans le domaine de l'apprentissage automatique pouvaient répondre à cette problématique en prédisant l'appartenance d'un patient à une classe donnée. Et dans un contexte d'afflux massif de patients aux urgences et de limitation d'accès au test de RT-PCR et à la TDM thoracique, cette option paraissait d'autant plus opportune. Nous avons donc mené cette étude pour prédire la présence de COVID-19 chez les patients lors de leur passage aux urgences afin de les orienter le plus précocement possible dans le bon circuit pour leur hospitalisation.
L'objectif de l'étude était d'exploiter l'ensemble des données clinico biologiques issues du dossier des patients pour améliorer la performance diagnostique des deux tests diagnostiques couramment utilisés : RT-PCR et TDM thoracique.
L'étude s'est concentrée sur l'utilisation des modèles d'apprentissage automatique suivants : régression logistique binaire, forêts aléatoires, et réseaux de neurones profonds pour analyser les données cliniques et biologiques des patients suspects d'être atteints de COVID-19 à l'admission aux urgences.
Nous avons recueilli un ensemble de données comprenant symptômes cliniques, antécédents médicaux, résultats de tests biologiques et d'imagerie dont nous avons extrait des variables prédictives pour entrainer et tester les modèles. Les performances des différents modèles ont été évaluées selon l'aire sous la courbe (AUC) Receiver Operating Characteristics (ROC).

Les résultats de notre étude ont montré que les modèles d'apprentissage automatique construits étaient capables de prédire le diagnostic de COVID-19 avec une précision élevée. Et plus exactement, le fait d'adjoindre l'apprentissage automatique a amélioré les performances des tests diagnostiques classiquement utilisés. Les performances des modèles étaient significativement meilleures que celles des méthodes diagnostiques traditionnelles : en intégrant les variables clinico biologiques dans un modèle de régression logistique, la performance diagnostique de l'imagerie thoracique est passée de 0,778 à 0,886 celle de la RT-PCR de 0,852 à 0,929. On a obtenu la même tendance avec le modèle de forêt aléatoire, la performance de l'imagerie thoracique passant à de 0,778 à 0,829 et celle de la RT-PCR de 0,852 à 0,903. Le réseau de neurones artificiels améliorait en valeur absolue la performance de l'imagerie, pas celle de la RT-PCR.
En définitive, le fait d'intégrer des variables clinico-biologiques dans des modèles d'apprentissage automatique permettait d'augmenter la performance diagnostique des deux méthodes diagnostiques traditionnelles et la meilleure performance était obtenue par le modèle de régression logistique qui combinait variables clinico biologiques et RT-PCR.

Donc le fait d'intégrer des variables clinico biologiques dans des modèles d'apprentissage automatique pouvait jouer un rôle crucial dans le diagnostic de la COVID-19, et fournir des résultats rapides et fiables en combinant les

données clinico biologiques avec les résultats de la RT-PCR et de la TDM thoracique. Cette approche paraissait très intéressante pour trier les patients suspects de COVID-19 dans un cadre où l'optimisation des ressources représentait un véritable enjeu et où il paraissait fondamental de réduire le nombre de faux négatifs dans le diagnostic de COVID-19 en cette période de pandémie. Les symptômes intégrés comme variables prédictives dans nos modèles étaient bien cohérents par rapport aux symptômes connus de la COVID-19, ce qui renforçait la validité des modèles développés.
En intégrant les avantages de l'apprentissage automatique dans le processus de diagnostic et de prise de décision clinique, les cliniciens pourraient bénéficier de modèles prédictifs basés sur l'ensemble des données du dossier patient en temps réel afin d'anticiper les évènements critiques des patients aux urgences. Cette approche permettrait une meilleure attribution des ressources avec une intervention plus précoce, contribuant ainsi à améliorer les résultats cliniques des patients.

Cependant, notre étude comporte également des limites.
Tout d'abord, la méthodologie repose sur des données rétrospectives qui sont issues du seul CHU de Rennes. Ceci pourrait limiter la généralisation des résultats à d'autres centres ou plus largement à d'autres contextes. Bien que les modèles aient montré des performances prometteuses, ils devraient être validés de manière prospective sur un plus grand échantillon de patients. En outre, la sélection des patients se basant sur certains symptômes choisis pour les définir comme suspects de COVID-19 a introduit un biais potentiel, car certains patients paucisymptomatiques ont pu ne pas être inclus dans l'effectif.
De plus, nous avons choisi de tester les coefficients de corrélation pour chaque paire de variable afin d'éliminer les variables redondantes. Cette méthode, assez classique, a pour avantage de ne pas reposer sur des éléments subjectifs. Le couplage avec un diagramme acyclique guidé (DAG) qui vise à représenter les relations causales entre variables en se basant sur l'expertise des cliniciens aurait pu permettre de diminuer encore le risque d'intégrer des variables redondantes. L'objectif de cet article, comme de l'ensemble de la thèse, était cependant la preuve de faisabilité de modèles prédictifs en situation d'urgences, ce qui n'était pas une donnée acquise, à priori. Dans la mesure où l'approche DAG implique une discussion entre cliniciens de différents horizons dont l'activité principale n'est pas centrée sur la recherche, en particulier en période de pandémie, son utilisation sera à envisager après obtention de la preuve de concept apportée par ce travail de thèse.
En fait, pour améliorer cette méthodologie, il serait important de valider les modèles sur des cohortes indépendantes provenant de différents centres hospitaliers. L'inclusion de patients paucisymptomatiques ou asymptomatiques pourrait enrichir la diversité des cas et renforcer la pertinence des modèles. De plus une comparaison directe entre les performances des modèles et les méthodes diagnostiques traditionnelles serait nécessaire pour évaluer leur véritable valeur ajoutée.

Finalement, cette étude offre des perspectives remarquables dans différents domaines qui pourraient bénéficier de ces avancées.
D'abord, dans l'amélioration des stratégies de tri et de gestion des ressources matérielles et humaines. L'utilisation de modèles d'apprentissage automatique pour le diagnostic de la COVID-19 aux urgences pourrait avoir un impact significatif en particulier dans les zones où les ressources sont limitées, c'est-à-dire dans la plupart des centres hospitaliers. En identifiant plus rapidement avec précision les cas positifs, les professionnels de santé pourraient mieux allouer les tests RT-PCR et les ressources médicales, en concentrant leur attention sur les patients les plus à risque. Cela pourrait réduire les temps d'attente, optimiser l'utilisation des lits d'hôpitaux et minimiser l'exposition des professionnels de santé aux cas non-urgents.

Ensuite, l'expérience de cette étude permet d'ouvrir la voie à l'intégration de l'intelligence artificielle dans les dispositifs médicaux utilisés dans les services d'urgence. Des outils de diagnostic automatisés, basés sur des modèles d'apprentissage automatique, pourraient être développés pour aider les professionnels de santé à prendre des décisions plus rapides et plus précises. Ces dispositifs pourraient être utilisés non seulement pour la COVID-19, mais aussi pour d'autres maladies infectieuses ou pour la prédiction de complications chez les patients initialement admis aux urgences.

Ces modèles pourraient aussi être développés pour suivre la propagation de la COVID-19 et d'autres maladies infectieuses en temps réel et donc être d'un apport considérable dans les systèmes de veille sanitaire. En analysant les données cliniques et épidémiologiques, ces modèles pourraient aider à détecter rapidement les zones à risque élevé, à prévoir les épidémies potentielles et à guider l'introduction de mesures de contrôle appropriées.

Enfin, l'intégration de l'intelligence artificielle dans les processus de diagnostic pourrait permettre une approche plus personnalisée des soins médicaux. Les modèles d'apprentissage automatique pourraient tenir compte des caractéristiques individuelles des patients, comme l'âge, les antécédents médicaux, les facteurs de risque et d'autres données cliniques, pour fournir des recommandations, de diagnostic et de traitement, plus adaptées à chaque cas. Cela pourrait améliorer l'efficacité des traitements et réduire les risques de complications.
Alors que ces avancées permettent de nombreuses opportunités, elles soulèvent également des questions éthiques importantes, notamment dans la confidentialité des données, la transparence des modèles, le consentement éclairé des patients et la prévention des biais algorithmiques. Il est crucial de garantir que ces technologies soient développées et déployées de manière éthique, en tenant compte des droits et des intérêts des patients.

En somme, notre étude, explorant l'utilisation de l'apprentissage automatique pour le diagnostic de COVID-19 aux urgences, offre des perspectives prometteuses pour améliorer la prise en charge des patients, optimiser les ressources médicales et contribuer à la surveillance épidémiologique. Cependant, une validation plus poussée, une diversification des données,

une intégration pratique et une réflexion éthique continue sont nécessaires pour exploiter le potentiel de ces technologies dans le domaine de la santé.

Maintenant, fort de ces conclusions, nous allons appliquer ces méthodes en les déclinant vers la prédiction des évènements critiques correspondant à des situations à risques de décès imminent. Nous allons donc dans la prochaine partie de cette thèse explorer à travers une dernière étude comment l'apprentissage automatique peut être utilisé pour anticiper ce besoin critique chez les patients admis aux urgences.

4. Partie III: Article 3 « Predicting critical care scenarios in hospitalized patients by using machine learning models »

Soumis dans Scientific Report

4.1. Résumé de l'article 3

Lors de situations critiques avec mise en jeu du pronostic vital, une prise de décision rapide représente un enjeu crucial. La possibilité de prédire précisément des évènements tels que la mortalité, l'intubation, la nécessité de réanimation cardiopulmonaire (RCP) ou le recours aux soins palliatifs peut aider les cliniciens, les patients et leurs familles à se préparer de manière adéquate. Au cours des dernières années, les modèles d'apprentissage automatique ont démontré leur intérêt dans la prédiction d'évènements rares et binaires, grâce à leur robustesse et à leur efficacité computationnelle. Notre étude a utilisé l'entrepôt de données « EHOP » du Centre Hospitalier Universitaire (CHU) de Rennes, en France, référentiel reconnu au niveau national de données de santé anonymisées. Notre objectif était de construire des modèles de forêts aléatoires à partir des données issues des visites des patients au Service des Urgences afin de prédire les quatre évènements en question chez les patients hospitalisés. Nous avons inclus tous les patients adultes admis après une visite aux urgences entre le 01/06/2019 et le 01/06/2022. Quatre modèles ont été développés pour prédire le décès l'intubation, la RCP et la décision de soins palliatifs. Les données utilisées pour extraire les variables prédictives étaient les antécédents médicaux, les traitements, les comptes-rendus de passage aux urgences, les constantes vitales, les comptes-rendus d'hospitalisation. 80 % de l'ensemble des données a été utilisé pour la construction du modèle et les 20 % restants pour son évaluation. Le critère de jugement principal était l'aire sous la courbe précision-rappel (PR-AUC) qui constitue une mesure robuste pour l'ensemble des données avec des distributions de classe déséquilibrées.
Les modèles ont été entrainés sur 31 982 patients et testés sur 7 996. Pour tous les modèles la PR-AUC était supérieure à 0,95 ce qui indique une performance exceptionnelle en matière de prédiction. La précision des modèles était toujours supérieure à 0,85, suggérant une grande capacité à éviter les faux positifs. En conclusion, nos modèles montrent une performance élevée dans la prédiction des résultats critiques tels que le décès, l'intubation, la RCP et les décisions de soins palliatifs pendant l'hospitalisation après une visite aux urgences. L'implémentation de ces modèles dans les systèmes d'aide à la décision clinique pourrait améliorer l'anticipation précoce des situations critiques chez les patients.

4.2. Article 3 : soumis dans Scientific Reports

Predicting critical care scenarios in hospitalized patients using machine learning models

Sonia Rafi, MD[1*]; Cedric Gangloff[1,2], MD, PhD; Ollivier Grimault[4], MD; Pauline Le Goff [5], MD; Ulysse Donval[5],MD; Guillaume Bouzillé, MD-PhD[3]; Marc Cuggia, MD-PhD[3]

*Corresponding author: sonia.rafi@univ-rennes1.fr
1 : Univ Rennes, INSERM, LTSI - UMR 1099, Rennes, France.
2 : Department of Emergency Medicine, Hôpital Privé Sévigné, Rennes, France.
3 : Search And Rescue Medical Unit, Department of Emergency Medicine, CHU de la Cavale Blanche, Brest, France
4 : Department of emergency CHU Rennes, France

Abstract
Rapid decision-making is crucial in critical care situations. Accurate prediction of outcomes, such as mortality, intubation, cardiopulmonary resuscitation (CPR) and need of palliative care, can help clinicians, patients and their families to prepare adequately. Our aim was to train random forest models on data from emergency department (ED) visits to predict these four outcomes in hospitalized patients. Data (medical history, treatments, vital signs, ED and hospitalization reports) were from the eHOP clinical data warehouse, Rennes University Hospital, France. All adult patients hospitalized after the ED visit from June 1, 2019, to June 1, 2022, were included. Four models were developed to predict death, intubation, CPR, and palliative care decisions. The primary metric used was the Area Under the Precision Recall Curve (PR-AUC), a robust measure for datasets with unbalanced class distributions. The models were trained and tested using data from 31,982 and 7,996 patients, respectively. In all models, the PR-AUC was ≥0.95, indicating exceptional prediction performance. The model precision was always >0.85, suggesting that they can avoid false positive predictions. Therefore, our models showed high performance for predicting these critical outcomes. Implementing these models in clinical decision-making support systems could improve the early anticipation of critical situations.

Background
The hospital admission of a patient can be fraught with critical events, leading to the patient's death[1]. This concerns 3 to 5% of all emergency admissions[2]. These critical care situations are a significant event for healthcare professionals/caregivers because they require ethical considerations before performing technical procedures that involve multiple participants in a limited time frame[3,4]. The aim of such ethical considerations is to determine whether the patient requires palliative care to avoid overzealous treatment when the

patient is beyond any therapeutic resource[5], or active care to compensate potential organ failure. In the second case, medical or surgical interventions must be performed as quickly as possible to maximize survival[6]. The two main interventions carried out to prevent death are intubation[7], to preserve the patient's ventilatory function, and cardiopulmonary resuscitation (CPR), to restore the circulatory function[8]. Therefore, critical care situations require rapid decision-making by healthcare teams in conjunction with the patient and family. Predicting death, the need of intubation, CPR, or palliative care would help all participants to make preparations and to anticipate the level of monitoring and care that should be provided to the patient.

In recent years, predictive machine learning models have shown their value in many fields, particularly in the healthcare sector [9-11]. Such models are based on variables to predict the occurrence of an event of interest, called the "outcome". There are several machine learning models, and the most frequently used are regression, neural networks, support vector machines, and random forests[12,13]. Random forests are considered very robust and computationally time-efficient models to predict the occurrence of rare and binary events [14,16]. Moreover, they do not require prior knowledge on the relationship between predictive variables and outcome because they are non-parametric algorithms.

The development of predictive machine learning models involves a training phase using an initial dataset, followed by a performance evaluation phase using different datasets. The dataset quality largely determines the model performance and its generalizability. Many data are generated during patient care: clinical imaging and blood testing results, antecedents, medical observations, discharge letters, hospital diagnoses. Their reuse is now facilitated by the existence of clinical data warehouses. Interestingly, many data are collected just before hospitalization when the patient is in the emergency department (ED). These data could be used as predictive variables for models to predict the occurrence of critical events in hospitalized patients. The aim of this study was to train random forests-based models using data collected during the ED visit to predict death, intubation, CPR, and palliative care in hospitalized patients.

Methods

Setting. Data on patients admitted to the adult ED of Rennes University Hospital, France, were retrospectively collected from the clinical data warehouse eHOP. The eHOP data warehouse, established in our university hospital for several years, is a French national reference in terms of storage and structuring, and uses deidentified health data[17-19].

Patient selection. The study included all adult patients (≥18 years of age) who were hospitalized after admission to the ED between June 1, 2019, and June 1, 2022.

Data collection.

The following data collected during the ED visit were extracted: medical history, ongoing treatments, visit reports, imaging reports, laboratory tests, vital signs, and treatments administered at the ED. Outcome data were

extracted from hospitalization reports and discharge diagnoses. Numerical data (e.g., blood pressure, heart rate, and respiratory rate) were extracted from structured HTLM tables within the comprehensive emergency room observations stored in the eHOP database. These observations encompassed a wide range of clinical scenarios. To handle the text-based information in the dataset, regular expressions were used, a versatile and widely used method for pattern recognition and text parsing. Regular expressions allowed us to systematically identify, capture, and structure relevant information from the unstructured text fields. This rigorous data extraction process ensured that both numerical and textual data were transformed into a suitable format for subsequent analysis and modeling, contributing to the robustness and comprehensiveness of our study.

Data pre-processing. A series of data pre-processing techniques was used to refine the data for modeling and analysis. **Text cleaning:** All non-alphanumeric characters were removed during data cleaning, and the text was normalized to eliminate punctuation and character casing variations. This approach allowed rationalizing the textual data for further processing. **Treatment of recurrent numerical variables:** When a numerical variable appeared recurrently in the dataset, only the first instance was retained for modeling. The aim of this step was to avoid redundancy and potential overfitting during modeling. **Data normalization:** Then, numerical data were normalized, a crucial step in which the variable values are adjusted to a common scale. This prevented potential inaccuracies in the estimation of variables in the model caused by scale differences[20]. **Missing data imputation:** The problem of missing data was addressed using an imputation method based on random forests[21]. This robust technique helped to maintain integrity when data points were missing. The test and train datasets were imputed separately. **Dimensionality reduction:** Dimensionality reduction was performed to simplify the model. This was achieved manually by categorizing treatments into broader groups. For example, all antihypertensive drugs were consolidated in a single variable that simplified the analysis without losing critical information[22].

Predicted variables.
Four models were created to predict the following events (one for each variable): death, intubation, CPR, and palliative care. The occurrence of these events of interest was identified using regular expressions employed in hospitalization reports and discharge diagnoses.

Selection of predicting variables. All 2013 clinical and laboratory variables present in the database were collected. However, only the 264 variables that appeared more than 1000 times in the dataset were selected for modeling. This threshold was arbitrarily set to eliminate very infrequent variables. Then, the correlation coefficient was calculated for each pair of variables and when it was >0.9, one of the variables was removed from the analysis. Finally, 168 variables were used for the analysis.

Data splitting.
The whole dataset was randomly split in two parts: the training dataset and the test dataset. The training dataset included 80% of the whole dataset and was used to build the models. The model performance was evaluated using the test dataset that included the remaining 20% of all data.

Class balancing. As the events predicted by the models were rare, the training dataset had to be rebalanced before modeling. Indeed, when the event of interest is rare in the training dataset, models tend to systematically predict its absence, resulting in an overall poor performance.
In this study, the ROSE approach was used[23]. This bootstrap-based technique enhances the performance of binary classification in rare classes by generating synthetic balanced samples and by simultaneously performing downsampling of the majority class. This methodological approach ensures a balanced representation of all classes, thus increasing the modeling predictive power and overall performance.

Model training. In this study, as outlined in Figure 1, a uniform methodology was followed for the development of the four random forest algorithms built with an ensemble of 1,000 decision trees. **Hyperparameter tuning:** It was achieved by constructing each random forest model using a set of 25 different hyperparameter combinations. These combinations included the total number of decision trees in the forest and the number of features considered by each tree during a split of nodes. **Cross-validation**: A k-fold cross-validation strategy[24] was used because it is instrumental for mitigating the risk of overfitting, a common pitfall in which machine learning algorithms can capture incidental statistical noise present in the training data. Overfitted models display high performance using training data, but falter in generalizing to unseen data. This makes them ineffective in providing appropriate predictions for new patients. K-fold cross-validation is a particularly robust method to circumvent this issue. The dataset was divided into k segments, or 'folds', and each model was trained using k-1 folds. The remaining fold was used for the model validation. This procedure was repeated k times (once per fold), thus necessitating additional computational resources. In our study, 10 folds were used for model construction. **Threshold tuning:** For each fold, the threshold used to align probabilities was systematically varied with class labels from 0 to 1, with increments of 0.1. This technique is particularly valuable for models designed to predict rare events, often associated with unbalanced datasets relative to the predicted variable. **Performance evaluation:** The models were constructed using a dedicated training dataset and their performance was evaluated using a distinct test dataset, not involved in the model training process. This procedure ensured unbiased performance measures by introducing the models to unseen data. The area under the Precision Recall Curve (PR-AUC) was used as the primary metric to assess the performance of each model. This metric is especially valuable in situations involving datasets that exhibit a class distribution imbalance for the predicted event because its analysis is primarily focused on the positive class [25,26]. Precision and Recall were calculated independently, as discrete

performance indicators for each model. Furthermore, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was quantified. This is a widely recognized metric used to assess and compare classifiers across a broad spectrum of applications, such as machine learning, biomedical research, and bioinformatics.

Variable importance. To evaluate and compare the relative significance of different variables within each model, the Mean Decrease in Impurity metric was used[27]. This metric computes the importance of the variable by gauging the total decrease in impurity of the node averaged over all trees in the ensemble. The mean impurity decrease is a representation of how much each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest model. Essentially, a higher average decrease in the impurity value for a given variable indicates that splitting the tree on this variable results in nodes with higher purity, thus signaling its importance in the classification process. This is a robust method to assess the contribution of each variable to the model predictive power.

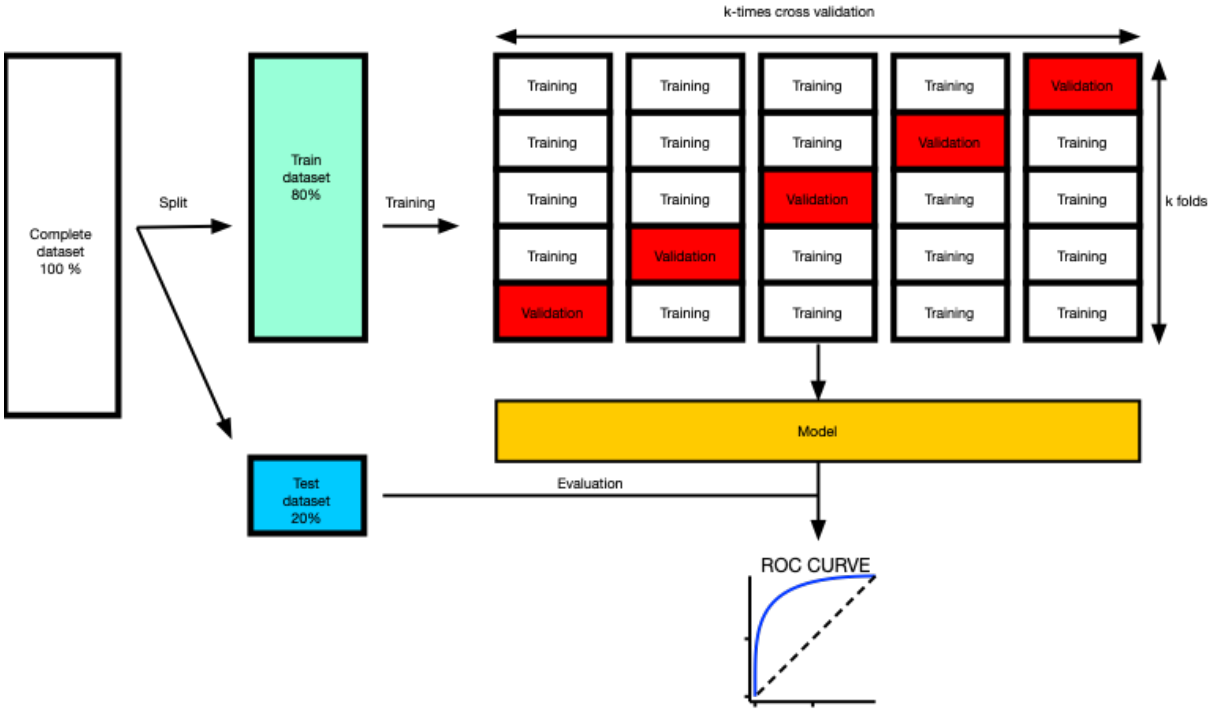

Figure 1. The methodology used for modeling was as follows: **1: Splitting.** The complete dataset was divided into a training dataset and a test dataset, with missing data imputed separately. **2: Training.** The training dataset, which included 80% of all data, was divided into 10 folds. Modeling was performed separately using these folds to produce models that performed as effectively as possible using new data, through a process of cross-validation, hyperparameter tuning and threshold tuning. **3: Performance measurement.** The models were evaluated using the remaining 20% of the complete dataset. This step used data not included in the training phase to obtain an unbiased evaluation of the model performance.

Software. Data extraction, processing, statistical analyses and model development were performed using R-Studio Server, version 1.4.1717, Rstudio PBC, 2009-2021. Specialized packages were employed for specific tasks: Dplyr, version 1.1.2, was used for data manipulation, and Purrr, version 1.0.1, facilitated data simplification. Regular expressions were executed with stringr, version 1.4.0. Oversampling and undersampling were managed with 3ROSE3, version 0.0-4. Models were constructed with tidymodels, version 0.1.3, and workflowsets, version 0.1.0. Cross-validation was done with Rsample, version 0.1.0. Hyperparameter tuning was executed with tune, version 0.1.6, and performance measurements were computed with parsnip, version 0.1.7.

Ethics

All procedure were carried out following the relevant guidelines and regulations. The permission to conduct research using data from the eHOP clinical data warehouse of Rennes University Hospital was granted by CNIL, the French National Commission for Informatics and Liberty. According to the French Data Protection Act of 6 January 1978, this study did not require informed written consent from the included participants because it only used information extracted from preexisting medical records and did not necessitate patient interaction or collection of identifiable private information. To ensure confidentiality, each data entry was deidentified. This study was approved by the Rennes University Hospital ethics committee (approval number: 23.88).

Results

Selection of patients.

In total, 195,330 adult patients were admitted to the ED between June 1, 2019, and June 1, 2022, and 39,977 patients were hospitalized (i.e. the complete dataset). The models were trained using 80% (n = 31,982) and evaluated using 20% (n=7,996) of the whole dataset.

The occurrence rates of the four predicted events in the patients included in the study are presented in Table 1.

Table 1. Occurrence of the four predicted events in the studied population

Outcome	True n (%)	False n (%)
Death	2239 (5.6%)	37738 (94.4%)
Intubation	758 (1.9%)	39219 (98.1%)
CPR	191(0.5%)	39786 (99.5%)
Palliative care	750 (1.9%)	39227(98.1%)

Predicting variables. In total, 168 clinical-laboratory variables were selected and considered as variables of interest for model building. The detailed list of

these variables is in Supplementary Table S1. The means and percentages for each outcome category were calculated for the numeric and binary categories, respectively.

Model performance. The performance metrics (AUC-ROC, precision, recall, and AUC-PR) for the four trained random forest models are presented in Table 2.

Predicted variable	Performance metric	Value
Death	AUC-PR	0.99
	Precision	1.00
	Recall	0.83
	AUC-ROC	0.99
	(F1-score	0.91)
Intubation	AUC-PR	0.95
	Precision	0.98
	Recall	0.71
	AUC-ROC	0.99
	(F1-score	0.82)
Cardiopulmonary ressuscitation	AUC-PR	0.96
	Precision	0.86
	Recall	0.99
	AUC-ROC	0.96
	(F1-score	0.92)
Palliative care	AUC-PR	0.99
	Precision	1
	Recall	0.30
	AUC-ROC	0.99
	F1-score	0.46

Table 2. Performance of the four models for predicting death, intubation, cardiopulmonary resuscitation, and palliative care. The following metrics were used: i) **Area under the Precision-Recall Curve** (AUC-PR), a combined measure of precision and recall; ii) **Precision** = TP / (TP + FP), it measures the accuracy of positive predictions made by the model; iii) **Recall** = TP / (TP + FN), it assesses the model ability to detect positive cases; and iv) **Area under the ROC curve** (AUC-ROC), it considers the model sensitivity and specificity. All metrics ranged from 0 to 1 (the minimum and maximum performance levels, respectively). TP = true positives, TN = true negatives, FP = false positives, FN = false negatives.

Importance of clinical-laboratory variables. The weights of the 20 most important variables for each model are presented in Figure 2. The complete set of weights for all variables is in Supplementary Table S2.

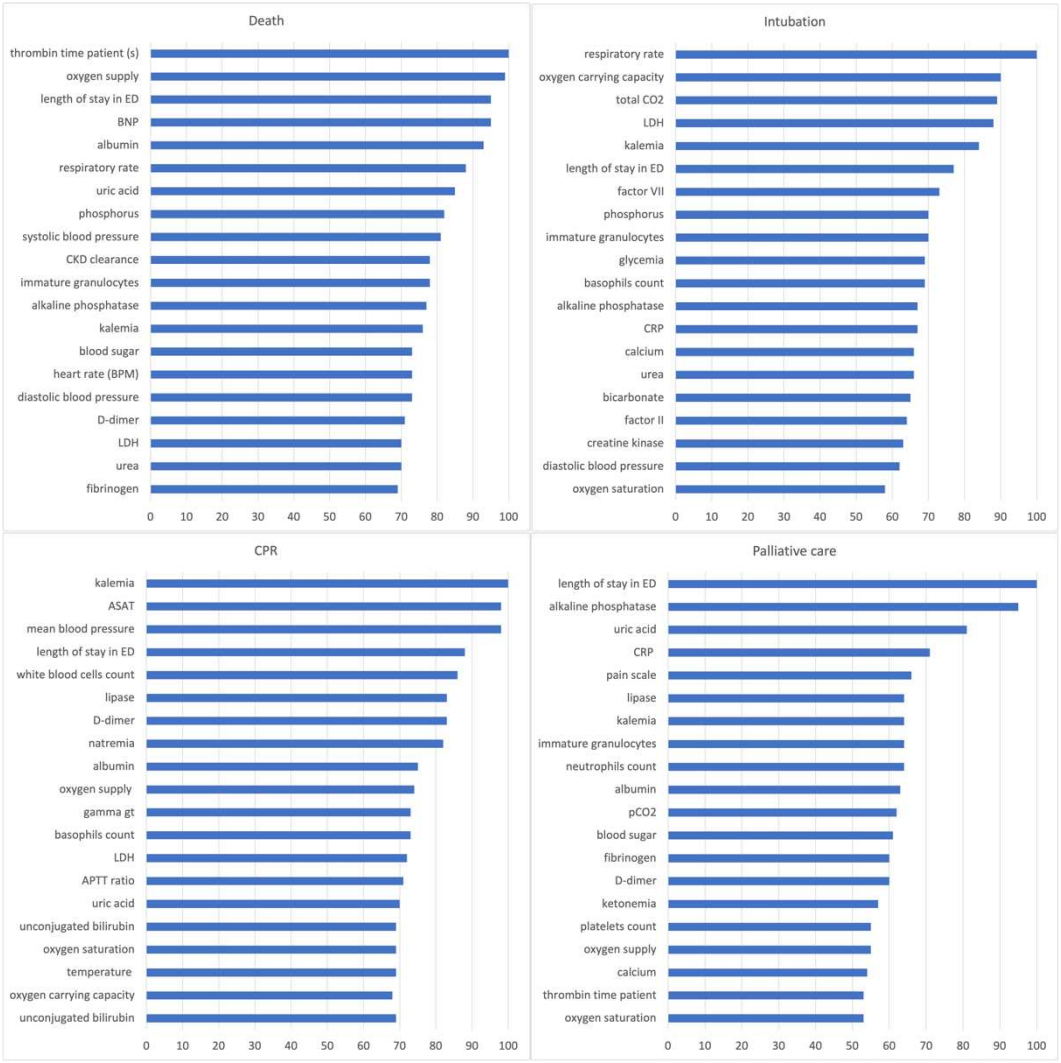

Figure 2. Ranking, in descending order, of the top 20 most important variables for each model. The bar length is proportional to the importance of the variable in the model and varies from 0 to 100. Change kalemia to blood K concentration; white blood cell count; basophil count; body temperature; D-dimers. Also, I would explain the abbreviations used in the figure.

Discussion.

As expected, the frequency of the predicted events was low (Table 1). It is more challenging to model a rare than a frequent event. However, the choice of random forest models and our overall methodology (class balancing and threshold tuning for predicted events) allowed the accurate prediction of rare events.

The overall performance of the models (one for each event) was excellent as indicated by the ROC-AUC and PR-AUC values (> 0.9 for all scenarios). These metrics are strong indicators of the model effectiveness. However, it is important to understand the distinct implications of these two metrics. The ROC curve is a graphical representation of sensitivity (or true positive rate) versus specificity (or false positive rate) using various thresholds. The ROC-AUC provides a single metric that summarizes this entire curve, effectively measuring the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. On the other hand, the PR curve plots the precision (or positive predictive value) against the recall (or sensitivity). The PR-AUC summarizes this curve and therefore, it is more sensitive to changes in the model performance on the positive class, especially when this is the minority class. It is a crucial metric when the goal is not just to discriminate positive from negative instances, but also to ensure that positive predictions are reliable and not diluted by false positive predictions. The fact that both ROC-AUC and PR-AUC values were >0.9 suggests the model strong discriminative power and also precision in predicting positive instances, even if rare. This model performance is highly advantageous because it ensures the robustness of the model predictive capacity and its reliability, even when faced with class imbalances typically found in real-world clinical data. Thus, the models might accurately predict the four events of interest, even in scenarios of rare occurrences. This is a promising result that must be confirmed using prospective data.

Furthermore, the precision of the four models was always >0.85. This indicates a strong capacity to avoid false positive predictions. Therefore, their implementation in medical software would not result in clinicians being overwhelmed by unnecessary alarms. This is an important parameter to consider when creating a cognitive aid system because fatigue caused by irrelevant alarms can lead to disregard alarms even when they are valid[28]. This issue has been extensively studied in the context of continuous patient monitoring in intensive care units and should be considered when developing decision-making support systems in this context.

The heterogeneity of the Recall values (sensitivity) needs to be discussed. The models to predict death and CPR exhibited high sensitivity (Recall >0.8), suggesting that they are reliable tools to identify patients who will die or need CPR during hospitalization. Conversely, the models for the prediction of intubation and for the prediction of palliative care displayed average (Recall = 0.71) and low (Recall = 0.30) sensitivity. This heterogeneity is a crucial issue when considering the future usage of these models. The models developed in this study are extremely reliable for predicting the occurrence of the four events of interest or the absence of CPR and death. Conversely, vigilance should not be lowered when the models suggest no imminent need of intubation or palliative care. Therefore, these models represent a promising tool to support clinical decision-making, but they should not be considered as a standalone software. Clinicians should see them as tools to support and inform their decision-making rather than a tool for ruling out patient management strategies.

The importance of predicting variables was consistent among the four models. These significant variables included temporal information (e.g., length of stay in ED), vital signs, and laboratory test values, thus providing a

multidimensional view of the patient's condition. The variable "time spent in the ED" ranked among the top six variables in all four models. This observation could be interpreted in several ways that all highlight the crucial role of time in patient management. First, it may signal that patients with a severe illness require thorough diagnoses and that this involves comprehensive testing, including clinical imaging and laboratory tests. These exams are time-consuming, thus extending the ED stay. In addition, discussions among different specialists to determine the optimal care plan could contribute to this longer stay. Second, some patients may present with multiple concurrent organ failures without immediate severity indicators. As these patients have a higher risk of mortality, secondary deterioration after admission is a major concern for clinicians, and this may warrant an extended observation period in the ED to ensure the patient's stabilization before hospitalization. Third, prolonged ED stay while waiting for hospital admission is considered an independent risk factor for morbidity and mortality, thus corroborating the significance of this variable in our study[29–31].

Moreover, the models emphasize the importance of vital signs in patient evaluation. Systolic blood pressure, oxygen saturation, heart rate, and respiratory rate emerged as critical predicting variables in the models. This finding is consistent with medical knowledge because vital signs offer a real-time snapshot of the patient's circulatory and respiratory state. They provide immediate feedback on the body ability to maintain essential physiological processes. These vital signs are particularly important because death is often preceded by respiratory and circulatory failure, thus providing potential opportunities for early intervention.

In our models, markers of organ failure also were important predicting variables. These include the Glasgow coma scale score, renal failure markers (e.g., creatinine clearance), cardiac markers (e.g. troponins and B-type natriuretic peptide), liver markers (e.g. aspartate transaminase and alkaline phosphatase) and the coagulation system (e.g. thrombin time, Quick time, and international normalized ratio). These markers provide valuable information on the functioning of key body systems. Their alteration/presence is the signature of critical conditions that can lead to death. Their consistent importance in all models corroborates their role as crucial indicators of the patients' health, especially in critically ill individuals.

Lastly, our models stress the importance of infection markers (e.g., body temperature, leukocyte count, and respiratory rate). This suggests that signs of infection play an important role in determining the risk and occurrence of the four events of interest. Infections can trigger systemic responses that disrupt the normal body functioning and increase the mortality risk. The automatic detection of a pattern involving infection markers with other variables identified in this study offers potential opportunities for the targeted monitoring of high-risk patients. This finding underscores the importance of considering the presence of infection as a significant factor that influences the health outcomes.

This study suggests that our models can be used to predict rare critical events in critical care; however, several limitations should be considered. First, this study was conducted in a single center, thus limiting the generalizability of the developed models to other healthcare facilities. However, the data were collected in the ED where standardized medical practices, commonly used in hospitals, were followed. Therefore, it is likely that our model will show similar performance when using multicenter data, although further validation is required to confirm this.

Second, our models displayed a robust predictive performance; however, it is important to emphasize that they are not intended to replace the clinical judgment. They should be viewed as tools that provide valuable insights based on factors observed during hospitalization. Clinicians should continue to exercise their expertise/judgement and consider predictions as supplementary information to guide their decision-making.

Third, the retrospective nature of this study means that the models were trained and evaluated using historical data. Prospective validation in real-time clinical settings is essential to determine their true utility and impact on patient outcomes. Furthermore, continuous monitoring and refinement of the models will be necessary because clinical practices evolve, and new variables or data sources become available.

Fourth, like with any predictive model, there is always the risk of bias or imbalance in the dataset. Despite our efforts to address class imbalances through class balancing and threshold tuning, real-word data may still present challenges related to data quality, missing values, and clinical practice variations. Our ongoing efforts to improve data quality and to capture additional variables will enhance the model accuracy and generalizability.

In conclusion, the machine learning models proposed in this study displayed an overall high performance for predicting death, intubation, CPR, or the need for palliative care during post-ED hospitalization. The incorporation of such models into medical software should be considered to improve clinical decision-making support systems and allow anticipating critical patient situations during hospital stays. Such proactive measures can contribute to the early ethical considerations and the timely deployment of technical interventions to manage organ failure. The capacity of these models to predict severe health outcomes is promising for improving patient management strategies.

References

1. Vallet, H. *et al.* Mortality of Older Patients Admitted to an ICU: A Systematic Review. *Crit. Care Med.* 49, 324–334 (2021).
2. Stewart, K., Choudry, M. I. & Buckingham, R. Learning from hospital mortality. *Clin. Med.* 16, 530–534 (2016).

3. Voumard, R. *et al.* Geriatric palliative care: a view of its concept, challenges and strategies. *BMC Geriatr.* 18, 220 (2018).
4. Ito, K. *et al.* Primary palliative care recommendations for critical care clinicians. *J. Intensive Care* 10, 20 (2022).
5. Lazris, A. Geriatric Palliative Care. *Prim. Care* 46, 447–459 (2019).
6. Kollef, M. H. *et al.* Timing of antibiotic therapy in the ICU. *Crit. Care* 25, 360 (2021).
7. Yamamoto, R. *et al.* Early intubation and decreased in-hospital mortality in patients with coronavirus disease 2019. *Crit. Care Lond. Engl.* 26, 124 (2022).
8. Yan, S. *et al.* The global survival rate among adult out-of-hospital cardiac arrest patients who received cardiopulmonary resuscitation: a systematic review and meta-analysis. *Crit. Care Lond. Engl.* 24, 61 (2020).
9. Zhang, Z., Ho, K. M. & Hong, Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. *Crit. Care Lond. Engl.* 23, 112 (2019).
10. Hashimoto, D. A., Witkowski, E., Gao, L., Meireles, O. & Rosman, G. Artificial Intelligence in Anesthesiology: Current Techniques, Clinical Applications, and Limitations. *Anesthesiology* 132, 379–394 (2020).
11. Blomberg, S. N. *et al.* Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation* 138, 322–329 (2019).
12. Zhang, Z., Cheng, S. & Solis-Lemus, C. Towards a robust out-of-the-box neural network model for genomic data. *BMC Bioinformatics* 23, 125 (2022).

13. Stenwig, E., Salvi, G., Rossi, P. S. & Skjærvold, N. K. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Med. Res. Methodol.* 22, 53 (2022).
14. Bai, Y., Huang, Z., Lam, H. & Zhao, D. Rare-event Simulation for Neural Network and Random Forest Predictors. *ACM Trans. Model. Comput. Simul.* 32, 18:1-18:33 (2022).
15. Bail, C. A. Lost in a random forest: Using Big Data to study rare events. *Big Data Soc.* 2, 2053951715604333 (2015).
16. Muchlinski, D., Siroky, D., He, J. & Kocher, M. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Polit. Anal.* 24, 87–103 (2016).
17. Delamarre, D., Bouzille, G., Dalleau, K., Courtel, D. & Cuggia, M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. *Stud. Health Technol. Inform.* 210, 702–706 (2015).
18. Lalanne, S. *et al.* Amoxicillin-Induced Neurotoxicity: Contribution of a Healthcare Data Warehouse to the Determination of a Toxic Concentration Threshold. *Antibiot. Basel Switz.* 12, 680 (2023).
19. Madec, J. *et al.* eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. *Stud. Health Technol. Inform.* 264, 1536–1537 (2019).
20. Borkin, D., Némethová, A., Michaľčonok, G. & Maiorov, K. Impact of Data Normalization on Classification Model Accuracy. *Res. Pap. Fac. Mater. Sci. Technol. Slovak Univ. Technol.* 27, 79–84 (2019).
21. Stekhoven, D. J. & Buhlmann, P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118 (2012).

22. Cunningham, P. University College Dublin.
23. Lunardon, N., Menardi, G. & Torelli, N. ROSE: a Package for Binary Imbalanced Learning. *R J.* 6, 79–89 (2014).
24. Berrar, D. Cross-Validation. in (2018). doi:10.1016/B978-0-12-809633-8.20349-X.
25. Sofaer, H., Hoeting, J. & Jarnevich, C. The area under the precision recall curve as a performance metric for rare binary events. *Methods Ecol. Evol.* 10, (2018).
26. Ozenne, B., Subtil, F. & Maucort-Boulch, D. The precision--recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J. Clin. Epidemiol.* 68, 855–859 (2015).
27. Han, H., Guo, X. & Yu, H. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. in *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* 219–224 (2016). doi:10.1109/ICSESS.2016.7883053.
28. Lewandowska, K. *et al.* Impact of Alarm Fatigue on the Work of Nurses in an Intensive Care Environment-A Systematic Review. *Int. J. Environ. Res. Public. Health* 17, 8409 (2020).
29. Groenland, C. N. L. *et al.* Emergency Department to ICU Time Is Associated With Hospital Mortality: A Registry Analysis of 14,788 Patients From Six University Hospitals in The Netherlands. *Crit. Care Med.* 47, 1564–1571 (2019).
30. Lin, S., Ge, S., He, W. & Zeng, M. Association of delayed time in the emergency department with the clinical outcomes for critically ill patients. *QJM Mon. J. Assoc. Physicians* 114, 311–317 (2021).

31. Stohl, S. *et al.* Impact of triage-to-admission time on patient outcome in European intensive care units: A prospective, multi-national study. *J. Crit. Care* 53, 11–17 (2019).

4.3. Résultats obtenus et discussion

Cet article propose une approche novatrice pour prédire les évènements critiques pouvant survenir chez les patients hospitalisés après une admission aux urgences. Nous avons utilisé un modèle de forêt aléatoire pour prédire quatre évènements critiques d'intérêt : le décès, l'intubation orotrachéale (IOT), la réanimation cardiopulmonaire et la décision d'arrêt et de limitation de thérapeutiques actives (LATA) ou nécessité de soins palliatifs. Pouvoir prédire ces évènements est particulièrement crucial pour les cliniciens, les patients ainsi que pour leurs proches puisque cela constitue une aide pour faire face de façon adéquate à des situations potentiellement létales et à des prises de décisions complexes. Les résultats de cette étude révèlent un potentiel significatif pour l'application de ces modèles prédictifs en médecine d'urgence.
Mais d'abord, il est important de noter que les quatre évènements d'intérêt : le décès, l'intubation, la réanimation cardiopulmonaire, la nécessité de soins palliatifs, sont des évènements rares si l'on considère leur incidence parmi l'ensemble des patients admis aux urgences. Il est donc particulièrement difficile de prédire ces évènements. Cependant, en utilisant un modèle de forêt aléatoire associé à une méthodologie globale incluant l'équilibrage des classes et l'ajustement des seuils pour les évènements prédits, nous avons réussi à prédire de manière précise ces évènements rares. Ceci souligne la robustesse de ces modèles et leur capacité à gérer les données médicales complexes et déséquilibrées.

Les modèles ont été entrainés et évalués sur une base de données de 39 977 patients adultes hospitalisés après admissions aux urgences du CHU de Rennes sur une période de trois ans.

Concernant les données textuelles, nous n'avons pas utilisé de traitement automatisé du langage (TAL) dans cet article. Les données textuelles étaient extraites par des expressions régulières (= REGEX).
Voici un exemple de REGEX pour déterminer la notion de céphalée : str_detect(crh_symptomes$texte, "(?<!pas de|pas|pas ete|sans|sans signe de|aucune|pas de notion de|sans notion de|aucune notion de|pas constate de|pas avoir de|pas eu de|pas presente de|ni) (migrai|cephal)")
Les documents analysés étaient les suivants :

- Textes des urgences : infirmiers et médecins
- Comptes rendus de sortie des services
- Comptes rendus de laboratoires
- Comptes rendus d'imagerie
- PMSI

Les notions recherchées étaient les suivantes :

- Comptes rendus des urgences :
  - Symptômes : malaise, suicide, intoxication, confusion, céphalée, traumatisme crânien, plaie, dyspnée, toux, fièvre, marbrures, asthénie, douleur thoracique, douleur abdominale, somnolence, œdèmes, anxiété, agitation, chute, fièvre
  - Orientation des patients : déchocage, hospitalisation
  - Examens complémentaires aux urgences : scanner, électroencéphalogramme, radiographie
  - Examens de laboratoire : PCR COVID
- Comptes rendus d'hospitalisation : dyspnée, intubation, limitation thérapeutique en service
- PMSI : décès, remplissage vasculaire, embolisation, transfusion, arrêt cardiaque, embolie pulmonaire, adrénaline, infarctus, radio interventionnelle, coronarographie

Certaines notions ont été extraites en quelques étapes simple (complexité intermédiaire) :

Par exemple, la notion d'intubation en service a été extraite en plusieurs étapes car notre modèle visait à détecter uniquement les intubations réalisées

1. Détecter intubation dans compte rendu urgences
2. Détecter intubation dans compte rendu d'hospitalisation
3. Un patient a été considéré intubé en service SI intubation en service positive ET intubation aux urgences négative

D'autres notions ont été beaucoup difficile à extraire (complexité élevée) : par exemple la détection de LATA :

##Étape 1: Créer une nouvelle variable avec les expressions contenant les mots clés.

```
keywords <- c("limitation", "lata", "paliative", "palliative", "palia", "pallia")
df0 <- df0 %>%
  mutate(texte_urg = str_replace_all(texte_urg, "\\s+", " "), # Replace line
breaks and consecutive spaces with a single space
         keyword_phrases = map(texte_urg, ~{
           keyword_phrases <- str_extract_all(.x,
sprintf("\\b(?:\\S+\\s+){0,10}%s(?:\\s+\\S+){0,10}\\b", keywords))#extraire les
expressions comprenant mots cles +/- 10 mots
           keyword_phrases <- unlist(keyword_phrases)
           keyword_phrases
         }))
```

##Étape 2: Tester chaque expression pour voir si elle contient un mot interdit.

```
exclusions <- c("pas","sans","membre","activite", "trauma","exam","mouve",
"refus", "artic","doul","ampli","ouver","mobili","effort","elective",
"perim","marche","flexi","exten","rota","abduc","addu")
```

```
df0 <- df0 %>%
  mutate(exclusion_present = map(keyword_phrases, ~{
    exclusion_present <- sapply(.x, function(phrase)
any(str_detect(tolower(phrase), exclusions)))
    exclusion_present
  }))
```

##Étape 3: Tester chaque case pour voir si elle contient au moins une expression avec mot clé sans mot interdit.

```
### var chr pour mot interdit
df0 <- df0 %>%
  mutate(exclusion_present_str = sapply(exclusion_present, paste, collapse
= ","))#on crée une variables string qui comporte des TRUE et des FALSE
(ou rien)

### test logique  : presence de FALSE dans "exclusions_present"
df0 <- df0 %>%
  mutate(lata_urgences = case_when(
    exclusion_present_str == "" ~ NA,
    grepl("FALSE", exclusion_present_str, ignore.case = TRUE) ~ TRUE,
    exclusion_present_str != "" & !grepl("FALSE", exclusion_present_str,
ignore.case = TRUE) ~ FALSE
  ))#si false détecté, =pas de mot interdit, si rien detecté, pas de mot clé
présent, si que des true détectés, mot clé présent mais avec mot interdit à
proximité


#Tester "pas d'acharnement thérapeutique"
df0 <- df0 %>%
   mutate(pas_acharnement_urg = str_detect(texte_urg, "pas acharnement")
| str_detect(texte_urg, "pas d’acharnement")| str_detect(texte_urg, "soins
confort")| str_detect(texte_urg, "soin confort"))

#Variable finale qui tient compte de lata et acharnement
df0$lata_urg <- ifelse(df0$lata_urgences == TRUE |
df0$pas_acharnement_urg == TRUE, TRUE, FALSE)
df1=df0%>%select(ID_SEJ,lata_urg)
df2=left_join(crh_var_supp,df1,by="ID_SEJ")
crh_var_supp=df2

rm(df0,df1,df2,keywords,exclusions)
```

Les performances des modèles ont été évaluées par l'aire sous la courbe ROC (ROC-AUC) et par l'aire sous la courbe précision-rappel (PR-AUC).
Ceci se justifie par le fait que s'agissant d'événement rares, la précision rappel était plus adaptée que le F1-score pour gérer les données déséquilibrées. En effet sur l'ensemble de l'effectif, la classe « décès » est beaucoup moins représentée que la classe « non-décès ».

Les performances globales de ces modèles étaient excellentes pour les quatre résultats étudiés, avec des valeurs supérieures à 0,9 pour ROC-AUC et PR-AUC. Ces mesures sont des indicateurs solides de l'efficacité des modèles, montrant leur capacité à discriminer avec précision les cas positifs des cas négatifs, même lorsque les cas positifs sont rares. Ceci suggère que ces modèles peuvent être utilisés comme outils de soutien pour les cliniciens, en fournissant des informations fiables pour guider leur prise de décision.
Un aspect important est la précision élevée des modèles, dépassant toujours 0,85. Cela signifie que les modèles ont une grande capacité à éviter les faux positifs. Cette caractéristique est cruciale dans un contexte médical où les fausses alertes peuvent être une source de stress et de fatigue pour les équipes de soins qui en retour risquent de négliger des alarmes légitimes. En utilisant ces modèles, les cliniciens peuvent être confiants dans le fait qu'ils ne seront pas submergés par des alarmes inutiles, ce qui permettra d'améliorer la qualité des soins.
Cependant, une limitation importante réside dans la disparité des performances entre les modèles pour les différents évènements. Les modèles ont montré une excellente sensibilité dans la prédiction du décès et de la réanimation cardiopulmonaire, mais une précision et une sensibilité plus faibles dans la prédiction de l'intubation et des soins palliatifs. Cela signifie que, bien que ces modèles soient des outils précieux pour certains évènements, ils peuvent ne pas être aussi fiables pour d'autres. Cette hétérogénéité des résultats souligne l'importance de la prudence lors de l'utilisation de ces modèles. Les cliniciens doivent vraiment considérer ces prédictions comme une aide ou une information supplémentaire pour guider leur prise de décision plutôt que comme des décisions autonomes.

Concernant les variables révélées comme importantes pour la prédiction de ces évènements, l'étude a mis en évidence plusieurs catégories cruciales : les variables temporelles, les variables de signes vitaux, les variables marqueurs d'insuffisance d'organe. Les variables temporelles, comme le temps passé aux urgences, étaient significatives dans tous les modèles, justifiant l'importance de la gestion du temps dans la prise en charge des patients. Les signes vitaux, tels que la pression artérielle systolique, la saturation en oxygène, la fréquence cardiaque et la fréquence respiratoire, sont essentiels, car ils fournissent un aperçu en temps réel de l'état circulatoire et respiratoire du patient. Enfin, les marqueurs d'insuffisance d'organes tels que le score de Glasgow, les marqueurs rénaux, cardiaques et hépatiques, ainsi que les marqueurs d'infection sont des indicateurs critiques de la santé du patient en particulier dans les cas graves, et ont également été déterminants dans la prédiction des résultats.

Cependant, malgré ces résultats prometteurs, plusieurs limites doivent être prises en compte.
Tout d'abord, cette étude a été menée dans un seul centre hospitalier, ce qui limite la généralisation des modèles à d'autres établissements de santé. Une validation à partir de plusieurs centres est nécessaire pour évaluer la portée réelle de ces modèles.
De plus, il est crucial de noter que ces modèles ne sont pas destinés à remplacer le jugement clinique mais plutôt à le soutenir. Les cliniciens doivent utiliser les prédictions comme des informations supplémentaires en tenant compte des spécificités de chaque cas. Ces modèles ne doivent pas être considérés comme des outils autonomes de décision médicale.
Ensuite, cette étude repose sur des données rétrospectives ce qui signifie qu'elle utilise des données historiques pour l'apprentissage et l'évaluation des modèles. Une validation prospective dans un environnement clinique en temps réel est nécessaire pour évaluer l'utilité de ces modèles et leur impact sur le devenir des patients.
Enfin, comme pour toute modélisation prédictive, il existe un risque de biais ou de déséquilibre dans les données. Malgré les efforts pour traiter ces problèmes, les données réelles présentent la difficulté liée à leur qualité, aux valeurs manquantes, aux variations de pratiques cliniques. Ainsi, le traitement des données réelles constitue un véritable défi. L'amélioration des procédés de gestion de qualité des données, ainsi que l'adjonction de variables complémentaires pourrait améliorer la précision et la possibilité de généraliser ces modèles.

En conclusion, cette étude met en évidence le potentiel des modèles d'apprentissage automatique pour la prédiction d'évènements critiques chez les patients hospitalisés après leur passage aux urgences. Ces modèles offrent des performances prometteuses, mais leur utilisation doit rester, en l'état, prudente et considérée comme une aide à la décision plutôt que comme une décision autonome. Des recherches futures devraient se concentrer une validation prospective dans des contextes cliniques réels et sur l'amélioration continue de la qualité des données pour garantir leur efficacité dans la pratique médicale.
Une fois que celles-ci auront été validées, la prochaine étape naturelle sera de concevoir une application qui intègre les modèles d'apprentissage automatique développés. Une telle application pourrait tirer parti des données en temps réel, collectées au moment de la visite aux urgences, pour fournir des informations instantanées aux professionnels de santé. Des solutions similaires ont déjà été développées avec succès dans d'autres domaines de la médecine comme la télémédecine et la surveillance à domicile. Par exemple des applications qui utilisent des données continues provenant de dispositifs portables pour surveiller les patients atteints de maladies chroniques ont montré des avantages significatifs en termes de gestion des soins. En intégrant ces modèles dans une application conviviale, les médecins aux urgences pourraient avoir accès à des prédictions en temps réel sur le pronostic vital du patient ce qui pourrait grandement contribuer à la prise de décision médicale dans des situations critiques.

5. Analyse globale des résultats

5.1. Comparaison des approches et résultats des trois articles

Ces trois articles contribuent à l'avancement des connaissances dans le domaine de la prédiction des évènements critiques aux urgences. Dans chaque article, nous avons abordé des problématiques spécifiques, et développé des approches adaptées pour y répondre et améliorer la prise en charge des patients. Nous allons comparer les approches et les résultats des trois études pour mettre en évidence leurs similitudes et leurs différences.

Le premier article se concentre sur la reconnaissance de l'arrêt cardiaque en contexte extrahospitalier par l'analyse des caractéristiques phonétiques du témoin qui appelle le centre 15. Les modèles d'apprentissage automatique, notamment celui utilisant les forêts aléatoires ont montré une capacité prometteuse à prédire cet événement à partir des caractéristiques vocales du discours du témoin avec une performance de 74,9 % (95 %IC [67,4 - 82,4]).
Le deuxième article vise à améliorer le diagnostic de la COVID-19 chez les patients hospitalisés durant la première vague de la pandémie en combinant données cliniques et biologiques recueillies dans le dossier médical patient. Les modèles de régression logistique, forêt aléatoire et réseau neuronal ont indiqué une amélioration potentielle des performances de la RT-PCR et de l'imagerie thoracique par l'utilisation de modèles d'apprentissage automatique intégrant les variables clinico biologiques dans ce contexte.
Le troisième article se penche sur la prédiction des évènements critiques qui correspondent à des scénarios où le risque de décès à brève échéance est important. La population étudiée correspond aux patients hospitalisés après un passage aux urgences et nous avons utilisé l'ensemble des données de routine du dossier médical informatisé. Les modèles de forêts aléatoires ont montré une performance exceptionnelle avec une PR-AUC supérieure à 0,95 pour la prédiction du décès, de l'intubation, de la réanimation cardiopulmonaire et de la décision de se limiter à des soins palliatifs.

Ces trois articles démontrent l'efficacité des modèles d'apprentissage automatique pour la prédiction d'évènements critiques en médecine d'urgence. Le dernier article (soumis) se distingue par ses performances particulièrement élevées, avec une PR-AUC supérieure à 0,95 indiquant une capacité à prédire ces évènements critiques. Par comparaison, les résultats du premier article, bien qu'encourageants, présentent des performances moindres, mais apportent des perspectives importantes pour la détection précoce de l'arrêt cardiaque extra hospitalier, qui plus est selon une approche très novatrice puisqu'elle utilise non pas directement les paramètres du patient lui-même, mais ceux du témoin. Ces paramètres sont pris comme marqueurs de stress de l'événement que l'on cherche à prédire. De la même façon, le deuxième article qui a pris naissance dans le contexte de la première vague de COVID-19 montre des performances prometteuses, mais encore à améliorer. C'est ce que nous avons fait dans notre dernier article où nous avons pu mettre à profit les méthodologies utilisées dans les deux précédents, en tirer des enseignements pour aboutir à ce dernier travail qui lui présente réellement des performances à la hauteur des objectifs visés dans le cadre de cette recherche.

Dans les deux premiers articles, nous avons utilisé les trois types de modèles d'apprentissage automatique : régression logistique binaire, forêt aléatoire, et réseaux de neurones artificiels. Alors que dans le dernier article, le seul modèle utilisé était la forêt aléatoire. Nous allons dans les prochaines lignes justifier cette démarche.
Notre objectif était, dans un premier temps, de tester la faisabilité de la prédiction d'évènements critiques d'intérêt par des méthodes d'apprentissage automatique et surtout de mettre en évidence ce que pouvait apporter le fait d'intégrer dans un modèle une variété importante de paramètres dans la prédiction des évènements.

Régression logistique d'un côté et random forest ou réseau de neurones de l'autre, ne sont pas des méthodes qui s'opposent, elles sont toutes des méthodes d'apprentissage automatique et nous les considérons comme un continuum de méthodes qui vont pouvoir traiter des données de plus en plus complexes.
En particulier, si l'on se penche sur la comparaison entre régression logistique et forêt aléatoire, elles donnent des performances équivalentes dans les deux premiers articles. Et en valeur absolue, la forêt aléatoire est supérieure à la régression logistique dans le premier article, et inférieure à la régression logistique dans le deuxième article. Nous pouvons alors faire les remarques suivantes.
Il faut d'abord rappeler que la régression logistique est un modèle qui va pouvoir inclure de nombreuses variables même si elles exercent une faible influence, et la valeur du coefficient correspondant à chaque variable permettra de connaître son influence sur la prédiction de l'événement d'intérêt. C'est un modèle simple et rapide à entrainer et facile d'interprétation. Cependant, la régression logistique va utiliser toutes les variables et donc beaucoup apprendre des données d'entrainement, ce qui va la rendre difficilement généralisable.
La forêt aléatoire permet d'utiliser plusieurs arbres de décision pour créer une forêt ce qui va améliorer le potentiel à généraliser le modèle. Chaque arbre de décision ne prend en compte qu'un sous-ensemble aléatoire des variables à chaque nœud de décision. Donc, si une variable n'a pas d'influence significative sur la prédiction, elle a moins de chance d'être choisie fréquemment pour diviser les arbres, ce qui va conduire à l'exclusion de certaines variables.
Cette notion est remarquable car quand on observe dans le 2ème article l'importance des variables selon le modèle (tableau 4), on constate qu'elles sont très différentes. Les variables cliniques ont une importance très faible dans le modèle de forêt aléatoire et ne vont donc être que très peu retenues. Ceci revêt un intérêt particulier dans la problématique de l'application de ces modèles en pratique clinique et leur acceptation par les cliniciens.
En somme, la forêt aléatoire va être plus complexe à interpréter que la régression logistique, moins intuitive, moins transparente car on ne visualisera pas toutes les variables, mais avec le potentiel de pouvoir traiter une quantité et une variété plus importante de données et d'obtenir de meilleures performances, avec des capacités de généralisation supérieures.
C'est à l'issue de ce raisonnement que nous n'avons retenu que ce modèle dans notre 3ème article. Et nous n'avons pas envisagé de le comparer, comme cela avait été fait dans les articles précédents avec un modèle de régression logistique car son utilisation aurait entrainé des temps de calcul ou de traitement

bien trop longs, de l'ordre de plusieurs semaines, par rapport à un modèle de forêt aléatoire, au vu du volume de données impliquées.

Enfin, au fur et à mesure de ces 3 articles une progression notable en termes de volume de données utilisé pour construire les modèles d'apprentissage automatique est observée. Dans le premier article, 820 patients et 18 variables ont été inclus dans l'étude ce qui représente déjà une taille d'échantillon significatif pour la reconnaissance des arrêts cardiaques extrahospitaliers. Dans le deuxième article les données de tous les adultes admis aux urgences pour suspicion COVID-19 entre le 20 mars 2020 et 5 mai 2020 ont été utilisées ce qui a considérablement augmenté la taille de l'échantillon par rapport au premier article, avec une cohorte de 5 196 patients, et 23 variables pour la construction des modèles. Enfin, dans le troisième article, une base de données de 39 977 patients avec 168 variables a été utilisée pour développer les modèles de prédiction pour des événements critiques aux urgences. Cette augmentation progressive du nombre de données traitées reflète l'évolution vers des analyses plus robustes et des modèles plus performants dans la prédiction de nos événements critiques d'intérêt.

5.2. Synthèse des principales contributions des articles à la prédiction des évènements critiques aux urgences

Dans cette section, nous allons approfondir notre analyse en présentant les principales contributions des trois articles à l'amélioration de la prédiction des évènements critiques aux urgences grâce aux techniques d'apprentissage automatique et en les mettant en perspective par rapport à l'état actuel des connaissances.

En effet, l'utilisation de l'apprentissage automatique dans le contexte de médecine d'urgence s'avère être une avancée majeure pour anticiper et gérer efficacement les situations critiques. Nos trois articles présentent des domaines d'application variés, mais convergent tous vers l'objectif commun d'améliorer la prise en charge des patients en contexte d'urgence, et ainsi d'en diminuer la morbimortalité.
Dans le premier article, nous nous sommes penchés sur la prédiction de l'arrêt cardiaque extrahospitalier par des méthodes d'apprentissage automatique basées sur les caractéristiques acoustiques de la voix de l'appelant, ceci s'inscrit dans le cadre plus large de la prédiction des évènements critiques des patients aux urgences.
Cette approche novatrice met en avant la possibilité de détecter cette situation particulièrement critique lors des appels au SAMU. Les techniques d'apprentissage automatique avaient été utilisées par Blomberg et al. dans ce contexte en utilisant les caractéristiques sémantiques du discours de l'appelant (47). Nous avons appliqué ces techniques en utilisant les caractéristiques acoustiques de la voix de l'appelant. L'utilisation de modèles d'apprentissage automatique, notamment la forêt aléatoire, a montré une performance encourageante avec une AUC à 74,9 %. Cette contribution est d'autant plus pertinente qu'à l'heure actuelle environ 30 % des arrêts cardiaques extra hospitaliers ne sont pas reconnus lors du premier appel (46). En combinant

l'analyse sémantique et acoustique, nous pourrions obtenir des performances optimales pour prédire cet événement particulièrement sensible.

L'année 2020 a été marquée par la survenue d'une pandémie mondiale ayant mis à mal l'ensemble de l'organisation des soins, et même du quotidien. Dans cette période particulière, nous avons appliqué notre méthodologie pour notre deuxième article. Avec cet article nous avons élargi le spectre en nous penchant sur l'amélioration du diagnostic de la COVID-19 chez les patients hospitalisés après leur admission aux urgences. Cette crise sanitaire a en effet exacerbé le besoin de diagnostic rapide et précis et notre article a exploré la façon dont l'apprentissage automatique pouvait se révéler comme un précieux atout dans ce contexte. L'enjeu était d'autant plus important que dans cette période-là, nous n'étions qu'au début de la crise sanitaire et la disponibilité du test RT-PCR était très restreinte avec des délais de rendu de résultats très longs. L'utilisation des données cliniques et biologiques extraites sur l'ensemble de la trajectoire du patient à partir du moment où il était admis aux urgences pour entraîner des modèles de régression logistique, de forêt aléatoire et de réseaux de neurones, a ouvert des perspectives d'amélioration de détection de cette pathologie. Nos modèles se sont, en effet, révélé être plus performants que lors de l'utilisation simplexe de l'imagerie thoracique ou du test RT-PCR pour ce diagnostic. Ces résultats sont essentiels dans un contexte de lutte contre la propagation du virus et l'assurance d'une prise en charge rapide et appropriée de cette maladie.

Le troisième article complète cette perspective en aboutissant sur la prédiction des évènements critiques qui constituent le décès ou les situations avec risque de décès à courte échéance : intubation, réanimation cardiopulmonaire ou décision de soins palliatifs. En effet, ces situations où le pronostic vital à court terme est en jeu, exigent des décisions rapides et précises. Nous avons surmonté la difficulté liée à la prédiction d'évènements rares concernant ces situations qui, heureusement, présentent des occurrences faibles en proportion de l'ensemble des patients admis aux urgences. Notre méthodologie globale avec utilisation de forêts aléatoires et aire sous la courbe précision-rappel nous a permis de prédire ces évènements malgré cette difficulté avec d'excellentes performances, en intégrant les données du patient sur l'ensemble de sa trajectoire à l'hôpital à partir du moment où il était admis aux urgences. L'efficacité de ces modèles suggère qu'ils peuvent être intégrés dans des systèmes d'aide à la décision clinique en permettant la détection précoce de ces évènements critiques.

Ces trois articles, bien que distincts dans leurs domaines d'application, partagent plusieurs éléments essentiels qui les rendent pertinents. Ils démontrent tous le potentiel puis, au fur et à mesure de l'avancement de nos travaux, l'efficacité de l'apprentissage automatique pour la prédiction d'évènements critiques et mettent en avant l'importance de l'analyse multimodale des données pour une prise de décision clinique améliorée.

En regardant de plus près ces contributions, nous pouvons identifier des thèmes communs et des leçons à retenir. Tout d'abord, l'intégration de données multiples est un élément-clé pour améliorer les performances des modèles d'apprentissage automatique. Et ceci, qu'il s'agisse de caractéristiques

phonétiques, de données cliniques ou biologiques, de données d'imagerie ou de données de routines, qu'elles soient structurées ou non.
De plus, les articles soulignent l'importance de la précision diagnostique et de la prédiction précoce pour une meilleure gestion des situations d'urgence ; dans le contexte du COVID-19, notamment lors des premières vagues, une détection rapide des cas positifs était impérative pour isoler les patients et ralentir la propagation du virus ; la reconnaissance des ACEH permet de débuter rapidement le massage cardiaque externe, ce qui peut être déterminant pour le pronostic du patient.
Un autre point que soulignent ces trois articles est l'accent mis sur la robustesse et la performance des modèles d'apprentissage automatique. Les forêts aléatoires se sont avérées particulièrement efficaces dans le premier et dans le troisième article. Cela suggère que ces modèles, précisément, peuvent constituer les choix les plus appropriés pour la prédiction d'évènements en contexte d'urgence.

Cependant, il est essentiel de noter que chaque article présente également ses propres limites et défis. Par exemple, le premier article donne une performance acceptable, mais loin d'être optimale. Celle-ci pourrait être améliorée par l'intégration des données vocales sémantiques et le modèle pourrait bénéficier d'une validation externe sur un ensemble de données plus vastes pour évaluer la possibilité de généraliser les résultats. De même le deuxième article souligne la nécessité de gérer les données manquantes et de lutter contre le surentraînement des modèles, qui constituent des défis courants dans le domaine de l'apprentissage automatique (52).
Enfin, le troisième article pose en particulier la question de l'interprétation des modèles, qui est cruciale pour assurer la confiance des cliniciens dans les décisions prises par les algorithmes.
Pour aller de l'avant, des recherches futures pourraient se concentrer sur l'intégration de ces approches dans la pratique clinique quotidienne. L'introduction de modèles d'apprentissage automatique dans les systèmes d'aide à la décision clinique pourrait être une étape importante pour améliorer la prise en charge des patients aux urgences. Cela nécessitera une étroite collaboration entre professionnels de la data et professionnels de santé pour garantir que les modèles soient adaptés aux besoins cliniques réels.

5.3. Perspectives d'amélioration et limites des méthodes d'apprentissage automatique

Les trois articles de cette thèse illustrent le fait que l'utilisation de l'apprentissage automatique pour la prédiction des évènements critiques aux urgences donne des résultats prometteurs. Cependant, pour maximiser son potentiel et son impact, plusieurs perspectives d'amélioration et des défis importants doivent être pris en compte.
Dans cette partie, nous allons aborder ce que pourraient apporter l'exploitation des données à partir des centres de données cliniques, les avancées de la recherche sur la qualité des données, l'élaboration d'applications en temps réel, et les défis liés à la fiabilité, la sécurité, l'efficacité et la performance des modèles (53).

Exploitation des données cliniques grâce aux centres de données cliniques
L'une des avancées clés pour l'amélioration des méthodes d'apprentissage automatique en médecine d'urgence et plus généralement en santé réside dans la possibilité d'exploiter les données provenant des centres de données cliniques. Les centres hospitaliers universitaires, tels que celui de Rennes sont devenus des pionniers dans la création de ces entrepôts de données cliniques, regroupant une vaste quantité d'informations médicales provenant de sources multiples comprenant entre autres les dossiers électroniques de santé, les résultats de laboratoire, les comptes-rendus d'imagerie et d'examens. Ces entrepôts représentent une mine d'informations pour la recherche en médecine d'urgence, offrant la possibilité d'explorer des ensembles de données volumineux et diversifiés pour la construction de modèles d'apprentissage automatique (54).
Cependant, l'exploitation de ces données nécessite une infrastructure informatique robuste, des outils performants d'analyse de données, et des protocoles stricts de protection de la vie privée et de sécurité des données. Les chercheurs doivent également collaborer étroitement avec les professionnels de santé pour comprendre la signification clinique des données et garantir leur pertinence dans la prise de décision médicale. C'est justement ce qui a été fait dans cette thèse et qui fait aussi l'originalité de ce travail, mais ce type de collaboration reste encore rare et gagnerait à être de plus en plus encouragé (55). En outre, la normalisation et la structuration des données dans les entrepôts de données cliniques sont des défis importants à relever pour garantir la qualité et la cohérence des informations utilisées pour l'apprentissage automatique.

Recherche centrée sur la qualité des données
La qualité des données est un élément crucial pour la fiabilité et la validité des modèles d'apprentissage automatique en médecine d'urgence. Les données cliniques sont souvent sujettes à des erreurs de saisie, des données manquantes, des incohérences et des biais. Par conséquent, la recherche en santé publique axée sur la qualité des données constitue un domaine essentiel sur lequel on gagnera à se référer dans nos travaux futurs (56). L'identification et la correction des erreurs de données, ainsi que la gestion des données manquantes sont des défis complexes mais incontournables pour garantir la précision et la fiabilité des modèles prédictifs (57).
Une approche prometteuse pour améliorer la qualité des données consiste à mettre en œuvre des stratégies automatisées pour la détection des anomalies et l'imputation des données manquantes. De plus, la création de normes de qualité des données et de protocoles de vérification de la qualité dans les centres de données cliniques peut contribuer à améliorer la fiabilité des informations utiles pour l'apprentissage automatique.

Caractère prospectif de la recherche
Une recherche prospective est nécessaire pour consolider les avancées réalisées dans le domaine de l'apprentissage automatique en médecine d'urgence. Cela implique la réalisation d'études cliniques visant à évaluer l'efficacité des modèles d'apprentissage automatique dans des contextes réels de prise en charge des patients. Les cliniciens doivent être impliqués dans la conception et la mise en œuvre de ces études pour garantir leur pertinence clinique.

De plus, la recherche prospective doit s'attaquer à des questions telles que l'acceptation des cliniciens de ces nouvelles technologies, la manière dont elles influencent les décisions médicales et leur impact sur le pronostic des patients. L'élaboration de directives cliniques basées sur des modèles d'apprentissage automatique est un domaine de recherche prometteur pour harmoniser la pratique médicale et garantir des normes élevées de soins aux patients.

Développement d'applications pour usage en temps réel

Une fois que la fiabilité, la rigueur et la sécurité des modèles d'apprentissage automatique auront été établies, le développement d'applications utilisables en temps réel par les cliniciens deviendra une priorité. Les outils d'intelligence artificielle doivent inspirer confiance et garantir sécurité, efficacité et performance pour être intégrés dans la pratique clinique quotidienne. En effet, les décisions prises dans ce domaine particulier touchent à la santé et à l'humain, ce qui constitue le caractère particulièrement sensible de la démarche.

Ces applications pourraient prendre la forme de systèmes d'aide à la décision clinique, d'outils de triage informatisé, ou même de dispositifs de surveillance en temps réel des patients aux urgences (58). L'objectif ultime est de fournir aux professionnels de santé des informations pertinentes et précises pour améliorer la prise en charge des patients, réduire les erreurs médicales et optimiser l'utilisation des ressources médicales et paramédicales.

Le développement d'applications en temps réel soulève des questions importantes en matière de réglementation, de responsabilité légale et de formation des cliniciens à l'utilisation de ces outils. Il est donc essentiel d'établir des protocoles clairs pour l'intégration de l'intelligence artificielle dans la pratique médicale en garantissant la conformité aux normes éthiques et légales en vigueur.

Rappelons qu'il existe désormais un cadre règlementaire inscrit dans le code de santé publique depuis août 2021, concernant l'obligation d'information au patient lors de l'utilisation d'outils d'intelligence artificielle, que ce soit à des fins diagnostiques ou de soins (40). C'est donc très récent et souligne bien le fait que nous ne sommes encore qu'au début de l'utilisation de ces méthodes. L'algorithme ne devra être considéré que comme aide à la décision, et son déploiement devra nécessairement être associé à une formation sur les limites de l'outil pour en assurer la confiance. Et ceci est indispensable pour apporter l'information adéquate au patient lors de sa prise en charge.

Et plus en amont, la participation des cliniciens au développement de ce type de méthodes en recherche en santé, implique un cadre règlementaire strict auquel ceux-ci sont déjà sensibilisés en termes d'éthique, mais il faut considérer, dans le cas précis de l'apprentissage automatique, du volume important de données à partager dans le temps entre différents intervenants.

Enseignement et sensibilisation

L'enseignement et la sensibilisation des professionnels de santé aux avantages et aux limites de l'apprentissage automatique en médecine d'urgence sont essentiels pour favoriser leur adoption. Les cliniciens devront être formés à l'utilisation de ces outils et être encouragés à les intégrer dans leur pratique de manière éthique et responsable (59). Des initiatives de sensibilisation telles que

des conférences, ateliers et des publications peuvent contribuer à informer la communauté médicale sur les développements récents en matière d'apprentissage automatique en médecine d'urgence. Les collaborations interdisciplinaires entre ingénieurs, professionnels de la data et cliniciens favoriseront un échange fructueux d'idées et de connaissances.

En outre, les codes et scripts informatiques élaborés pour la construction des modèles d'apprentissage automatique dans le cadre de nos recherches sont susceptibles d'être mis à disposition de la communauté scientifique, permettant ainsi leur réutilisation, éventuellement après adaptation spécifique, pour d'autres applications de prédictions d'évènements en santé, contribuant ainsi à la diffusion et à l'essor de l'intelligence artificielle dans le domaine médical.

L'objectif ultime de notre projet de recherche est, à terme, d'optimiser le parcours de soin.
En revanche, il est indispensable de disposer d'outils performants et validés pour atteindre cet objectif. Notre thèse constitue la première étape du projet, puisque le point de départ est un état des lieux où l'on constate que ces outils ne sont pas développés.
Il s'agit d'abord de les construire, les évaluer, puis les confronter à la clinique et refaire un feed-back par rapport à ce retour des cliniciens pour finalement développer les outils qui permettront réellement d'aboutir à l'optimisation du parcours de soins.
Donc dans la suite de notre travail, l'étape naturelle serait la réalisation d'études multicentriques, prospectives, en ayant résolu l'ensemble des problèmes touchant à la qualité des données, tout en respectant les règles d'éthique et de confidentialité des données.
Les systèmes apprenants ne peuvent donc pas encore remplacer le travail clinique, tant que l'ensemble de ces étapes n'aura pas été réalisé.

En définitive, notre travail représente une démonstration de faisabilité puisque ces méthodologies sont encore très novatrices dans la pratique clinique notamment pour la prise de décision et en contexte de soins critiques.

6. Conclusion

L'utilisation de modèles d'apprentissage automatique pour la prédiction des évènements critiques chez les patients admis aux urgences ouvre la voie à une transformation profonde de la médecine moderne. En élargissant notre analyse et en examinant ces résultats dans le contexte plus vaste de la santé, nous pouvons entrevoir un avenir où les avancées technologiques transforment radicalement la manière dont diagnostiquons, traitons et prévenons les maladies. Un domaine prometteur qui accompagne cette recherche est le développement des centres de données cliniques (CDC), tel que celui du CHU de Rennes avec son entrepôt de données eHOP qui a été utilisé dans ce travail pour construire nos modèles. Les CDC constituent une source inestimable d'informations médicales issues de divers supports, comme l'imagerie, les comptes-rendus médicaux de consultations ou d'examens, les pancartes paramédicales, les notes de cliniciens, les données de laboratoire, et plus encore. Ils offrent une vision complète du parcours du patient et peuvent être exploités pour identifier des tendances, des corrélations et des facteurs de risque. En utilisant ces données, des modèles peuvent être développés pour prédire non seulement les évènements critiques, mais aussi les maladies chroniques telles que le diabète, les pathologies cardiovasculaires, certains cancers, de manière plus précoce que ce qui n'est fait actuellement, en gardant en tête que la prévention va constituer aujourd'hui la clé de voute du succès des systèmes de santé.

Le défi du « data reuse » dont notre travail constitue une véritable approche, représente une opportunité substantielle. Il s'agit de l'utilisation de toutes les données de la trajectoire du patient hospitalisé, quelle que soit leur forme. Cela inclut non seulement les données structurées, telles que les résultats de tests, mais aussi les données textuelles, telles que les rapports médicaux, ou simplement les notes d'observation clinique. Le Traitement Automatique du Langage (TAL) joue un rôle clé ici. Les modèles de TAL peuvent par exemple extraire des informations précieuses à partir de notes médicales non structurées.

Cependant pour que le « data re-use » deviennent une réalité exploitable, il est essentiel de résoudre le problème de la qualité des données. Actuellement, de nombreuses données médicales sont hétérogènes prises à l'état brut, ce qui les rend difficiles à utiliser pour construire des modèles fiables d'apprentissage automatique. Des efforts considérables sont développés dans la recherche pour transformer ces tas de données non normalisées en données de qualité. Ceci implique, au-delà du processus d'extraction, de nettoyer les données, les structurer et de les normaliser afin de pouvoir facilement les utiliser pour la modélisation. Dans notre travail, nous avons suivi ces étapes méthodologiques, cependant, il serait bénéfique à l'avenir d'adopter une approche automatisée dédiée à la gestion de la qualité des données. Cette démarche constitue un domaine de recherche distinct sur lequel nous pourrons capitaliser dans nos futurs travaux de recherche.

L'utilisation de ces technologies soulève des questions importantes d'interprétabilité et de fiabilité. Par exemple, dans le cas du TAL, il est essentiel de développer des méthodes pour expliquer comment les modèles ont pris leurs

décisions. Cela garantit que les cliniciens comprennent les raisons derrière les recommandations des modèles, sans les ressentir comme des « boîtes noires ».

Dans la même ligne, l'acceptabilité clinique est un enjeu clé : les modèles prédictifs doivent être conçus pour être des outils d'aide à la décision, et non pas pour être des substituts des cliniciens. Les alarmes générées par ces modèles doivent être pertinentes et fiables, mais elles ne doivent pas non plus être une source de stress pour les professionnels de santé. Il est impératif de maintenir un équilibre pour que ces technologies soient véritablement bénéfiques dans la pratique clinique.

Enfin, le respect rigoureux de la confidentialité des données demeure la préoccupation centrale et c'est précisément à ce niveau que les méthodes d'anonymisation doivent continuer de se développer. Cela permet de garantir que les données puissent être partagées en toute sécurité entre les chercheurs et les institutions, favorisant ainsi la collaboration et les avancées dans le domaine des données massives en santé.

En conclusion, notre travail marque le début d'une ère passionnante dans la médecine prédictive. Ceci s'inscrit parfaitement dans le cadre de la médecine 5P : personnalisée, préventive, prédictive, participative et pertinente. Cette ambition peut être atteinte en poursuivant les collaborations entre professionnels et en développant l'interopérabilité des systèmes comme cela a été réalisé dans cette thèse. En exploitant pleinement les données cliniques, en garantissant l'interprétabilité et l'acceptabilité clinique, nous pouvons créer un avenir où les soins de santé seront plus efficaces et plus sûrs.

7. Références bibliographiques

(1) Raita, Y. et al, Emergency department triage prediction of clinical outcomes using machine learning models. Crit. Care 23, 64 (2019)

(2) Rajpurkar P. et al, Chexnet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv: 1711.05225v3 Dec 2017 https://stanfordmlgroup. github.io/projects/chexnet

(3) Deo RC. Machine Learning in Medicine. Circulation. 2015 Nov 17;132(20):1920-30.

(4) Wissenberg M, Lippert FK, Folke F, et al. Association of National Initiatives to Improve Cardiac Arrest Management With Rates of Bystrander Intervention and Patient Survival After Out-Of-Hospital Cardiac Arrest. JAMA. 2 oct 2013;310(13):1377

(5) Hasselqvist-Ax I, Riva G, Herlitz J et al. Early Cardiopulmonary Resuscitation in Out-Of-Hospital Cardiac Arrest. New Engl J Med. 11 juin 2015; 372(24):2307-15

(6) Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, Gurka D, Kumar A, Cheang M. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Crit Care Med. 2006 Jun ;34(6):1589-96.

(7) Godinjak A, Iglica A, Rama A, Tančica I, Jusufović S, Ajanović A, Kukuljac A. Predictive value of SAPS II and APACHE II scoring systems for patient outcome in a medical intensive care unit. Acta Med Acad. 2016 Nov;45(2):97-103.

(8) Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG, Sirio CA, Murphy DJ, Lotring T, Damiano A, et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. Chest. 1991 Dec;100(6):1619-36. doi: 10.1378/chest.100.6.1619. PMID: 1959406.

(9) Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA. 1993 Dec 22-29;270(24):2957-63. doi: 10.1001/jama.270.24.2957. Erratum in: JAMA 1994 May 4;271(17):1321. PMID: 8254858.

(10) Metnitz PG, Moreno RP, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR; SAPS 3 Investigators. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. Intensive Care Med. 2005 Oct;31(10):1336-44. doi: 10.1007/s00134-005-2762-6. Epub 2005 Aug 17. PMID: 16132893; PMCID: PMC1315314.

(11) Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Crit Care Med. 2006 May;34(5):1297-310. doi: 10.1097/01.CCM.0000215112.84523.F0. PMID: 16540951.

(12) Moreno RP, Metnitz PG, Almeida E, Jordan B, Bauer P, Campos RA, Iapichino G, Edbrooke D, Capuzzo M, Le Gall JR; SAPS 3 Investigators. SAPS 3--From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med. 2005 Oct;31(10):1345-55. doi:

10.1007/s00134-005-2763-5. Epub 2005 Aug 17. Erratum in: Intensive Care Med. 2006 May;32(5):796. PMID: 16132892; PMCID: PMC1315315.
(13) Moreno R, Apolone G. Impact of different customization strategies in the performance of a general severity score. Crit Care Med. 1997 Dec;25(12):2001-8. doi: 10.1097/00003246-199712000-00017. PMID: 9403750.
(14) Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B, Gouzes C, Lepage E, Moine P, Villers D. Mortality prediction using SAPS II: an update for French intensive care units. Crit Care. 2005;9(6):R645-52. doi: 10.1186/cc3821. Epub 2005 Oct 6. PMID: 16280063; PMCID: PMC1414016.
(15) Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009 Jun 4;338:b606. doi: 10.1136/bmj.b606. PMID: 19502216.
(16) Teres D, Lemeshow S. Why severity models should be used with caution. Crit Care Clin. 1994 Jan;10(1):93-110; discussion 111-5. PMID: 8118735.
(17) Moons KG, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009 Jun 4;338:b606. doi: 10.1136/bmj.b606. PMID: 19502216.
(18) van Galen LS, Dijkstra CC, Ludikhuize J, Kramer MH, Nanayakkara PW. A Protocolised Once a Day Modified Early Warning Score (MEWS) Measurement Is an Appropriate Screening Tool for Major Adverse Events in a General Hospital Population. PLoS One. 2016 Aug 5;11(8).
(19) Fernández-Delgado M, Sirsat MS, Cernadas E, Alawadi S, Barro S, Febrero-Bande M. An extensive experimental survey of regression methods. Neural Netw. 2019 Mar;111:11-34. doi: 10.1016/j.neunet.2018.12.010. Epub 2018 Dec 21. PMID: 30654138.
(20) Li D, Zhang H, Liu Z, Huang J, Wang T. [Deep residual convolutional neural network for recognition of electrocardiogram signal arrhythmias]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi. 2019 Apr 25;36(2):189-198.
(21) Moody GB, Mark RG. A Database to support Development and Evaluation of Intelligent Intensive Care Monitoring. Computers in Cardiology 23:657-660 (1996)
(22) Goldberger, A., Amaral,L., Glass,L., Hausdorff, J., Ivanov, P.C., Mark, R.; … & Stanley, H.E. (2000). Physiobank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 101 (23), pp; e215-e220
(23) Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit Care Med. 2011 May;39(5):952-60. doi: 10.1097/CCM.0b013e31820a92c6. PMID: 21283005; PMCID: PMC3124312.
(24) Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016 May 24;3:160035. doi: 10.1038/sdata.2016.35. PMID: 27219127; PMCID: PMC4878278.

(25) Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman LH, Celi LA, Mark RG. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023 Jan 3;10(1):1. doi: 10.1038/s41597-022-01899-x. Erratum in: Sci Data. 2023 Jan 16;10(1):31. Erratum in: Sci Data. 2023 Apr 18;10(1):219. PMID: 36596836; PMCID: PMC9810617.
(26) Escobar GJ, LaGuardia JC, Turk BJ, Ragins A, Kipnis P, Draper D. Early detection of impending physiologic deterioration among patients who are not in intensive care: development of predictive models using data from an automated electronic medical record. J Hosp Med. 2012 May-Jun;7(5):388-95.
(27) Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Crit Care Med. 2018 Apr;46(4):547-553.
(28) Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ, Das R. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. JMIR Med Inform. 2016 Sep 30;4(3):e28.
(29) Desautels T, Das R, Calvert J, Trivedi M, Summers C, Wales DJ, Ercole A. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. BMJ Open. 2017 Sep 15;7(9):e017199.
(30) Gajic O, Malinchoc M, Comfere TB, Harris MR, Achouiti A, Yilmaz M, Schultz MJ, Hubmayr RD, Afessa B, Farmer JC. The Stability and Workload Index for Transfer score predicts unplanned intensive care unit patient readmission: initial development and validation. Crit Care Med. 2008 Mar;36(3):676-82.
(31) Calvert J, Saber N, Hoffman J, Das R. Machine-Learning-Based Laboratory Developed Test for the Diagnosis of Sepsis in High-Risk Patients. Diagnostics (Basel). 2019 Feb 13;9(1):20.
(32) Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med. 2016 Feb;44(2):368-74.
(33) Layeghian Javan S, Sepehri MM, Aghajani H. Toward analyzing and synthesizing previous research in early prediction of cardiac arrest using machine learning based on a multi-layered integrative framework. J Biomed Inform. 2018 Dec;88:70-89.
(34) Li D, Zhang H, Liu Z, Huang J, Wang T. [Deep residual convolutional neural network for recognition of electrocardiogram signal arrhythmias]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi. 2019 Apr 25;36(2):189-198.
(35) de Souza Filho EM, Fernandes FA, Wiefels C, de Carvalho LND, Dos Santos TF, Dos Santos AASMD, Mesquita ET, Seixas FL, Chow BJW, Mesquita CT, Gismondi. RA. Machine Learning Algorithms to Distinguish Myocardial Perfusion SPECT Polar Maps. Front Cardiovasc Med. 2021 Nov 11;8:741667.
(36) Johnson KR, Hagadorn JI, Sink DW. Alarm Safety and Alarm Fatigue. Clin Perinatol. 2017 Sep;44(3):713-728. doi: 10.1016/j.clp.2017.05.005. Epub 2017 Jul 14. PMID: 28802348.

(37) Wilken M, Hüske-Kraus D, Röhrig R. Alarm Fatigue: Using Alarm Data from a Patient Data Monitoring System on an Intensive Care Unit to Improve the Alarm Management. Stud Health Technol Inform. 2019 Sep 3;267:273-281. doi: 10.3233/SHTI190838. PMID: 31483282.
(38) Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, Cadotte DW. The impact of machine learning on patient care: A systematic review. Artif Intell Med. 2020 Mar;103:101785. doi: 10.1016/j.artmed.2019.101785. Epub 2019 Dec 31. PMID: 32143792.
(39) Azencott CA. Machine learning and genomics: precision medicine versus patient privacy. Philos Trans A Math Phys Eng Sci. 2018 Sep 13;376(2128):20170350. doi: 10.1098/rsta.2017.0350. PMID: 30082298.
(40) LOI n° 2021-1017 du 2 août 2021 relative à la bioéthique
(41) Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny ML, Delamarre D, Raimbert V, Lemordant P, Cuggia M. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. Stud Health Technol Inform. 2019 Aug 21;264:1536-1537.
(42) Gräsner JT, Lefering R, Koster RW, Masterson S, Böttiger BW, Herlitz J, Wnent J, Tjelmeland IB, Ortiz FR, Maurer H, Baubin M, Mols P, Hadžibegović I, Ioannides M, Škulec R, Wissenberg M, Salo A, Hubert H, Nikolaou NI, Lóczi G, Svavarsdóttir H, Semeraro F, Wright PJ, Clarens C, Pijls R, Cebula G, Correia VG, Cimpoesu D, Raffay V, Trenkler S, Markota A, Strömsöe A, Burkart R, Perkins GD, Bossaert LL; EuReCa ONE Collaborators. EuReCa ONE-27 Nations, ONE Europe, ONE Registry: A prospective one month analysis of out-of-hospital cardiac arrest outcomes in 27 countries in Europe. Resuscitation. 2016 Aug;105:188-95.
(43) Bossaert LL, Perkins GD, Askitopoulou H, Raffay VI, Greif R, Haywood KL, Mentzelopoulos SD, Nolan JP, Van de Voorde P, Xanthos TT; ethics of resuscitation and end-of-life decisions section Collaborators. European Resuscitation Council Guidelines for Resuscitation 2015: Section 11. The ethics of resuscitation and end-of-life decisions. Resuscitation. 2015 Oct;95:302-11.
(44) Hirlekar G, Jonsson M, Karlsson T, Bäck M, Rawshani A, Hollenberg J, Albertsson P, Herlitz J. Comorbidity and bystander cardiopulmonary resuscitation in out-of-hospital cardiac arrest. Heart. 2020 Jul;106(14):1087-1093.
(45) Nolan JP, Sandroni C, Böttiger BW, Cariou A, Cronberg T, Friberg H, Genbrugge C, Haywood K, Lilja G, Moulaert VRM, Nikolaou N, Olasveengen TM, Skrifvars MB, Taccone F, Soar J. European Resuscitation Council and European Society of Intensive Care Medicine guidelines 2021: post-resuscitation care. Intensive Care Med. 2021 Apr;47(4):369-421.
(46) Duhem H, Viglino D, Debaty G. Future prospects in out-of-hospital cardiac arrest management. Méd Intensive Réa 30(4)297-310. 2020 May.
(47) Blomberg SN, Folke F, Ersboll AK, Christensen HC, Torp-Pedersen C, Sayre MR, Counts CR, Lippert FK. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. Resuscitation. 2019 May; 138:322-329.
(48) Mendoza E, Carballo G. Acoustic analysis of induced vocal stress by means of cognitive workload tasks. J Voice. 1998 Sep;12(3):263-73.

(49) Lu H, Stratton CW, Tang YW. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. J Med Virol. 2020 Apr;92(4):401-402.
(50) Novel coronavirus (COVID-19) situation [internet]. Available from: https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd
(51) Mizumoto K, Chowell G. Estimating Risk for Death from 2019 Novel Coronavirus Disease, China, January–February 2020. Emerg Infect Dis [Internet]. 2020 Jun [cited 2020 Mar 22];26(6).
(52) Kernbach JM, Staartjes VE. Foundations of Machine Learning-Based Clinical. Prediction Modeling: Part II-Generalization and Overfitting. Acta Neurochir Suppl. 2022;134:15-21.
(53) Safran C. Update on Data Reuse in Health Care. Yearb Med Inform. 2017 Aug;26(1):24-27. doi: 10.15265/IY-2017-013. Epub 2017 Sep 11.
(54) Delamarre, D., Bouzille, G., Dalleau, K., Courtel, D. & Cuggia, M. Semantic integration of medication data into the EHOP Clinical Data Warehouse. *Stud. Health Technol. Inform.* 210, 702–706 (2015).
(55) Scott IA. Demystifying machine learning: a primer for physicians. Intern Med J. 2021 Sep;51(9):1388-1400.
(56) Stausberg J, Harkener S. Data Quality and Data Quantity: Complements or Contradictions? Stud Health Technol Inform. 2023 Jun 29;305:24-27.
(57) Yang HS, Rhoads DD, Sepulveda J, Zang C, Chadburn A, Wang F. Building the Model. Arch Pathol Lab Med. 2023 Jul 1;147(7):826-836.
(58) Corbin CK, Maclay R, Acharya A, Mony S, Punnathanam S, Thapa R, Kotecha N, Shah NH, Chen JH. DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record. J Am Med Inform Assoc. 2023 Aug 18;30(9):1532-1542.
(59) Al-Edresee T. Physician Acceptance of Machine Learning for Diagnostic Purposes: Caution, Bumpy Road Ahead! Stud Health Technol Inform. 2022 Jun 29;295:83-86.

8. Annexes

8.1. Tableaux supplémentaires article 3

Supplementary table 1 : The 168 variables used for model building.
Means and proportions were calculated in the true and false groups for each of the 4 outcomes. p-values for the difference between groups were also calculated.

Category	Variable	Death false	Death true	Death p-value	Intubation false	Intubation true	Intubation p-value	CPR false	CPR true	CPR p-value	Palliative care false	Palliative care true	Palliative care p-value
Patient history	stroke	0.11	0.14	<0.001	0.11	6.20%	<0.001	11%	12%	0.8	0.11	0.13	0.085
Patient history	hypertension	0.3	0.38	<0.001	0.3	0.21	<0.001	30%	33%	0.7	0.3	0.35	0.019
Patient history	heart attack	6.60%	0.1	<0.001	6.90%	5.00%	0.13	6.80%	12%	0.038	6.80%	9.70%	0.007
Patient history	arrhythmia	7.30%	0.11	<0.001	7.60%	4.00%	<0.001	7.60%	3.70%	0.13	7.50%	0.11	<0.001
Patient history	COPD	4.60%	7.30%	<0.001	4.70%	6.50%	0.084	4.80%	6.30%	0.6	4.70%	6.50%	0.071
Patient history	diabete	0.13	0.16	0.002	0.14	0.11	0.2	14%	15%	0.8	0.14	0.13	>0.9
Patient history	addiction	9.30%	9.60%	0.8	9.30%	0.11	0.2	9.30%	9.40%	>0.9	9.40%	6.50%	0.031
Patient history	depression	0.11	0.11	0.8	0.11	0.11	>0.9	11%	7.90%	0.4	0.11	0.1	0.8
Patient history	ulcer	5.90%	7.60%	0.004	6.00%	4.60%	0.3	6.00%	4.20%	0.6	6.00%	7.70%	0.13
Patient history	cancer	0.19	0.26	<0.001	0.2	0.13	<0.001	20%	16%	0.5	0.19	0.37	<0.001
Usual treatments	Anti-hypertensive	0.2	0.26	<0.001	0.2	0.14	<0.001	20%	18%	0.7	0.2	0.22	0.5
Usual treatments	diuretic	0.12	0.2	<0.001	0.12	7.50%	<0.001	12%	14%	0.9	0.12	0.15	0.15
Usual treatments	Anti-thrombotic	0.25	0.31	<0.001	0.25	0.17	<0.001	25%	26%	>0.9	0.25	0.3	0.012
Usual treatments	anti-ulcer	0.13	0.17	<0.001	0.13	9.20%	0.003	13%	16%	0.5	0.13	0.18	0.002
Usual treatments	lipid lowering	0.11	0.13	0.051	0.11	7.30%	0.003	11%	13%	0.7	0.11	0.11	>0.9
Usual treatments	psychotropic	0.13	0.14	0.3	0.13	6.30%	<0.001	13%	9.90%	0.5	0.13	0.16	0.027
Usual treatments	painkillers	0.11	0.13	0.002	0.11	5.40%	<0.001	11%	7.30%	0.3	0.11	0.16	<0.001
Reasons for admission	malaise	9.80%	8.00%	0.023	9.70%	5.90%	0.002	9.70%	12%	0.7	9.70%	8.80%	0.7
Reasons for admission	confusion	8.00%	0.1	0.004	8.20%	5.10%	0.009	8.20%	4.20%	0.13	8.10%	0.11	0.004
Reasons for admission	headache	8.10%	3.10%	<0.001	7.80%	5.40%	0.047	7.80%	4.20%	0.2	7.90%	3.20%	<0.001
Reasons for admission	dyspnea	0.2	0.35	<0.001	0.2	0.31	<0.001	20%	26%	0.14	0.2	0.31	<0.001
Reasons for admission	cough	9.50%	0.12	<0.001	9.50%	0.16	<0.001	9.70%	9.90%	>0.9	9.70%	9.70%	>0.9
Reasons for admission	fever	0.17	0.15	0.066	0.16	0.16	0.8	17%	9.90%	0.051	0.17	0.15	0.4
Reasons for admission	weakness	0.19	0.29	<0.001	0.2	0.15	0.006	20%	17%	0.6	0.2	0.35	<0.001
Reasons for admission	chest pain	0.11	0.1	0.5	0.11	9.50%	0.3	11%	14%	0.6	0.11	9.50%	0.3
Reasons for admission	abdominal pain	0.11	7.50%	<0.001	0.11	4.90%	<0.001	11%	4.70%	0.029	0.11	8.70%	0.2
Reasons for admission	fall	0.19	0.18	0.6	0.19	9.10%	<0.001	19%	13%	0.1	0.19	0.17	0.7
Reasons for admission	dyspnea	0.28	0.46	<0.001	0.29	0.43	<0.001	29%	36%	0.1	0.29	0.42	<0.001
Vital signs	systolic blood pressure (mmHg)	142	135	<0.001	141	132	<0.001	141	137	0.2	141	136	<0.001
Vital signs	diastolic blood pressure (mmHg)	82	78	<0.001	81	79	0.011	81	77	0.042	81	78	<0.001
Vital signs	mean blood pressure (mmHg)	101	97	<0.001	101	96	<0.001	101	97	0.084	101	98	<0.001
Vital signs	heart rate (BPM)	87	90	<0.001	87	96	<0.001	87	90	0.2	87	91	<0.001

Vital signs	temperature (°C)	36.95	36.82	<0.001	36.94	37.09	0.046	36.95	36.76	0.11	36.95	36.88	0.2
Vital signs	oxygen saturation (%)	96.54	95.32	<0.001	96.5	94.78	<0.001	96.47	95.42	0.024	96.48	95.73	<0.001
Vital signs	glasgow coma scale	14.83	14.23	<0.001	14.81	13.85	<0.001	14.8	15	<0.001	14.8	14.79	0.8
Vital signs	respiratory rate (breaths per minute)	23	26	<0.001	23	26	<0.001	23	23	0.7	23	25	0.043
Vital signs	pain scale	3	1.8	<0.001	2.9	3.5	0.5	3	2	0.5	2.9	2.8	0.8
Vital signs	capillary hemoglobin(g/dl)	12.1	9.5	0.076	12	8.4	0.027	12	11.2	0.6	12	8.5	0.016
Vital signs	blood sugar (mg/dl)	127	135	0.017	127	135	0.2	127	148	0.11	127	123	0.4
Vital signs	ketonemia (mmol/l)	1.54	1.01	0.091	1.51	0.82	0.087	1.5	1.66	0.9	1.5	1.14	0.5
Symptoms reported by doctors	confusion	0.12	0.17	<0.001	0.13	8.40%	0.002	13%	9.40%	0.4	0.12	0.19	<0.001
Symptoms reported by doctors	headache	9.80%	4.30%	<0.001	9.60%	7.40%	0.13	9.50%	5.80%	0.2	9.60%	4.40%	<0.001
Symptoms reported by doctors	skin wound	8.70%	6.70%	0.004	8.70%	5.80%	0.021	8.60%	8.40%	>0.9	8.70%	4.50%	<0.001
Symptoms reported by doctors	cough	0.14	0.18	<0.001	0.14	0.21	<0.001	14%	16%	0.7	0.14	0.19	0.002
Symptoms reported by doctors	fever	0.18	0.17	0.4	0.18	0.19	0.9	18%	15%	0.4	0.18	0.18	>0.9
Symptoms reported by doctors	mottled skin	5.20%	0.14	<0.001	5.70%	9.50%	<0.001	5.70%	8.90%	0.2	5.70%	0.11	<0.001
Symptoms reported by doctors	weakness	0.21	0.3	<0.001	0.22	0.16	0.001	22%	20%	0.8	0.21	0.39	<0.001
Symptoms reported by doctors	chest pain	0.27	0.25	0.2	0.27	0.21	0.001	27%	30%	0.7	0.27	0.25	0.6
Symptoms reported by doctors	abdominal pain	0.11	8.10%	<0.001	0.11	4.40%	<0.001	11%	3.70%	0.004	0.11	9.10%	0.2
Symptoms reported by doctors	fall	0.19	0.18	0.2	0.19	0.1	<0.001	19%	13%	0.11	0.19	0.16	0.12
Complementary tests	cerebral scan	0.1	0.12	0.079	0.1	9.00%	0.5	10%	10%	>0.9	0.1	0.13	0.13
Complementary tests	chest scan	7.10%	0.12	<0.001	7.20%	0.17	<0.001	7.30%	15%	<0.001	7.40%	7.30%	>0.9
Complementary tests	abdominal scan	7.40%	7.00%	0.8	7.40%	4.10%	0.002	7.40%	3.70%	0.15	7.40%	8.00%	0.8
Complementary tests	electrocardiogram	0.45	0.5	<0.001	0.45	0.43	0.5	45%	42%	0.8	0.45	0.45	>0.9
Complementary tests	chest x-ray	0.19	0.28	<0.001	0.19	0.21	0.5	19%	22%	0.6	0.19	0.28	<0.001
Complementary tests	pelvis x-ray	7.60%	6.90%	0.5	7.70%	4.40%	0.003	7.60%	3.70%	0.12	7.60%	8.00%	>0.9
Complementary tests	COVID-19 PCR	0.13	0.14	0.094	0.13	9.90%	0.054	13%	8.90%	0.3	0.13	0.15	0.088
Prescriptions	activated clotting time	0.5	0.48	0.3	0.5	0.41	<0.001	50%	38%	0.005	0.5	0.42	<0.001
Prescriptions	pain monitoring	0.23	0.2	<0.001	0.23	0.13	<0.001	23%	13%	0.003	0.23	0.18	0.007
Prescriptions	temperature monitoring	0.26	0.28	0.1	0.26	0.25	0.9	26%	19%	0.14	0.26	0.22	0.059
Prescriptions	IV line mentionned	9.00%	0.16	<0.001	9.20%	0.22	<0.001	9.40%	14%	0.14	9.40%	8.80%	0.8
Prescriptions	vital signs monitoring	0.41	0.4	0.8	0.41	0.34	<0.001	41%	31%	0.035	0.41	0.36	0.044
Prescriptions	neurological status monitoring	0.24	0.26	0.1	0.24	0.25	>0.9	24%	21%	0.6	0.24	0.23	0.9
Prescriptions	catheter	0.21	0.24	<0.001	0.21	0.24	0.2	21%	19%	0.8	0.21	0.19	0.6
Prescriptions	urine strip	0.12	0.1	0.11	0.12	5.90%	<0.001	12%	8.40%	0.4	0.12	0.11	0.9

Prescriptions	administration of usual treatments	0.5	0.48	0.4	0.5	0.35	<0.001	50%	39%	0.012	0.5	0.45	0.063
Prescriptions	IV rehydration	0.22	0.27	<0.001	0.22	0.24	0.4	22%	19%	0.5	0.22	0.21	0.8
Prescriptions	aerosol	5.40%	9.60%	<0.001	5.60%	0.1	<0.001	5.70%	6.80%	0.8	5.70%	6.10%	0.9
Prescriptions	oral hydration	8.00%	8.30%	>0.9	8.10%	6.10%	0.13	8.00%	6.80%	0.8	8.00%	8.50%	0.9
Prescriptions	painkiller	0.17	0.12	<0.001	0.17	8.60%	<0.001	17%	6.30%	<0.001	0.17	0.14	0.1
Prescriptions	antihypertensive	7.70%	9.20%	0.042	7.90%	4.10%	<0.001	7.80%	7.30%	>0.9	7.80%	6.70%	0.5
Prescriptions	diuretic	7.70%	0.15	<0.001	8.20%	4.60%	0.002	8.10%	12%	0.14	8.10%	9.50%	0.4
Prescriptions	antithrombotic	9.00%	9.80%	0.4	9.10%	5.00%	<0.001	9.00%	5.80%	0.3	9.00%	9.10%	>0.9
Prescriptions	antidiabetic	0.1	0.11	0.6	0.1	9.10%	0.6	10%	10%	>0.9	0.1	9.70%	>0.9
Prescriptions	antibiotic	0.1	0.15	<0.001	0.11	0.12	0.3	11%	9.40%	0.9	0.11	0.12	0.3
Prescriptions	o2 (l/min)	4.7	6.1	<0.001	4.8	6.3	<0.001	4.9	5.1	0.8	4.8	5.4	0.2
Blood tests	red blood cells count (million/mm3)	4.2	3.98	<0.001	4.19	4.21	0.6	4.19	4.17	0.8	4.2	3.87	<0.001
Blood tests	platelets count (cells/mm3)	248	246	0.6	248	240	0.2	248	242	0.5	248	255	0.3
Blood tests	white blood cells count (million/mm3)	10.3	12.3	<0.001	10.4	10.7	0.3	10.4	12	0.007	10.4	12.6	<0.001
Blood tests	neutrophils count (million/mm3)	75	80	<0.001	75	78	<0.001	75	78	0.03	75	80	<0.001
Blood tests	eosinophils count (million/mm3)	1.19	0.78	<0.001	1.18	0.78	<0.001	1.17	0.86	0.022	1.18	0.77	<0.001
Blood tests	basophils count (million/mm3)	0.55	0.45	<0.001	0.54	0.47	<0.001	0.54	0.54	0.9	0.54	0.44	<0.001
Blood tests	mean corpuscular RBC volume (fl)	92	94	<0.001	92	93	0.002	92	93	0.4	92	93	0.002
Blood tests	mean corpuscular platelets volume (fl)	8.59	8.76	<0.001	8.59	8.67	0.2	8.59	8.92	0.006	8.59	8.63	0.5
Blood tests	mean corpuscular hemoglobin (g/dl)	30.86	31.14	<0.001	30.87	31.22	0.051	30.88	30.83	0.9	30.88	30.86	0.9
Blood tests	mean corpuscular hemoglobin concentration (g/dL)	33.6	33.27	<0.001	33.58	33.44	0.031	33.58	33.28	0.016	33.59	33.16	<0.001
Blood tests	red cell distribution width (%)	14.89	15.99	<0.001	14.95	15.09	0.3	14.95	15.46	0.046	14.92	16.35	<0.001
Blood tests	quick test patient (s)	16.1	20	<0.001	16.3	16.4	0.9	16.3	18.1	0.2	16.3	18.2	0.004
Blood tests	prothrombin (%)	86	72	<0.001	86	84	0.14	86	77	0.007	86	77	<0.001
Blood tests	international normalized ratio (no unit)	1.25	1.56	<0.001	1.26	1.26	>0.9	1.26	1.41	0.15	1.26	1.4	0.001
Blood tests	activated partial thromboplastin time (s)	35	39	<0.001	36	36	0.2	36	37	0.2	36	37	<0.001
Blood tests	APTT ratio (no unit)	1.04	1.13	<0.001	1.05	1.07	0.2	1.05	1.1	0.14	1.05	1.1	<0.001
Blood tests	D-dimer (mg/dl)	2.6	4	0.001	2.7	3	0.6	2.7	4	0.5	2.7	3.2	0.4
Blood tests	fibrinogen (g/l)	4.71	4.25	0.034	4.64	5.06	0.4	4.65	5.64	0.4	4.65	4.78	0.7
Blood tests	immature granulocytes (million/l)	3.15	3.58	0.4	3.23	2.37	0.042	3.22	2.47	0.3	3.11	4.76	0.1
Blood tests	factor V (%)	79	74	0.1	78	73	0.4	78	71	0.5	78	85	0.13
Blood tests	thrombin time patient (s)	17.37	19.26	0.085	17.49	18.35	0.7	17.51	16	0.4	17.4	22.04	0.13

Blood tests	Vitamin B9 (nmol/L)	32.7	33.5	0.3	32.7	34.4	0.14	32.7	36.8	0.2	32.8	32.5	0.8
Blood tests	factor II (%)	67	61	0.04	67	70	0.4	67	80	0.2	67	70	0.4
Blood tests	factor VII (%)	52	46	0.045	51	54	0.7	51	83	0.1	51	53	0.6
Blood tests	factor X (%)	64	59	0.11	64	73	0.11	64	74	0.02	63	72	0.046
Blood tests	natremia (mmol/l)	137.4	136.3	<0.001	137.4	136.5	0.002	137.4	135.5	0.002	137.4	136.7	0.006
Blood tests	kalemia (mmol/l)	4.12	4.27	<0.001	4.13	4.17	0.4	4.13	4.37	0.019	4.13	4.25	<0.001
Blood tests	chloride (mmol/l)	99.6	97.8	<0.001	99.6	98.1	<0.001	99.6	96.6	<0.001	99.6	98.6	<0.001
Blood tests	bicarbonate (mmol/l)	23	22	<0.001	23	20.6	<0.001	23	20.6	0.01	23	22.8	0.7
Blood tests	glycemia (mmol/l)	7.47	8.19	<0.001	7.49	8.45	0.002	7.5	8.78	0.026	7.5	7.77	0.13
Blood tests	urea (mmol/l)	7.6	12	<0.001	7.9	8.8	0.014	7.9	10	0.013	7.8	10.8	<0.001
Blood tests	uric acid (µmol/l)	356	431	<0.001	360	415	0.007	361	436	0.033	360	388	0.082
Blood tests	protidemia (g/l)	72	70	<0.001	71	71	0.7	71	71	0.4	71	69	<0.001
Blood tests	calcium (mmol/l)	2.33	2.3	<0.001	2.33	2.26	<0.001	2.33	2.29	0.11	2.33	2.32	0.7
Blood tests	phosphorus (mmol/l)	1.04	1.22	<0.001	1.05	1.31	<0.001	1.05	1.33	0.044	1.05	1.14	0.008
Blood tests	CRP mg/l	55	85	<0.001	56	90	<0.001	57	86	0.026	56	93	<0.001
Blood tests	BNP (pg/ml)	3.572	7.786	<0.001	3.948	4.45	0.7	3,932	8,501	0.1	3.923	5.394	0.1
Blood tests	troponin (ng/l)	70	228	<0.001	81	71	0.6	80	181	0.2	80	115	0.3
Blood tests	CKD clearance (ml/min)	105	100	<0.001	105	107	0.2	105	103	0.5	105	103	0.3
Blood tests	creatine kinase (IU/l)	685	716	0.8	679	1.039	0.2	687	833	0.7	689	568	0.4
Blood tests	ASAT (IU/L)	69	116	<0.001	71	87	0.1	71	64	0.7	71	92	0.1
Blood tests	ALAT (IU/L)	56	75	0.007	57	55	0.7	57	50	0.7	57	71	0.3
Blood tests	alkaline phosphatase (IU/L)	109	174	<0.001	113	104	0.2	113	119	0.7	110	226	<0.001
Blood tests	gamma gt (IU/L)	126	199	<0.001	131	127	0.8	131	120	0.7	128	240	<0.001
Blood tests	lipase (IU/L)	204	211	>0.9	197	693	0.13	204	49	<0.001	206	81	<0.001
Blood tests	pH (no units)	7.43	7.43	0.14	7.43	7.38	<0.001	7.43	7.41	0.2	7.43	7.43	>0.9
Blood tests	pCO2 (mmHg)	39	40	0.3	39	42	0.2	39	37	0.4	39	41	0.4
Blood tests	total co2 (mmol/L)	27.2	27.3	0.9	27.2	26.1	0.3	27.2	24.6	0.018	27.2	28.2	0.2
Blood tests	pO2 (mmHg)	87	92	0.1	87	91	0.3	87	78	0.2	87	86	0.8
Blood tests	oxygen saturation (%)	92.2	92.5	0.6	92.2	93.4	0.051	92.3	91.1	0.6	92.3	92.2	>0.9
Blood tests	hemoglobin (g/dl)	12.76	12.28	0.004	12.7	13.22	0.2	12.71	12.88	0.8	12.74	11.6	<0.001
Blood tests	oxygen carrying capacity (ml/dl)	16.6	16.1	0.02	16.5	17.4	0.062	16.6	16.6	>0.9	16.6	15.1	0.001
Blood tests	base excess (mEq/l)	3.41	4.49	<0.001	3.49	4.5	0.075	3.51	2.51	0.11	3.48	4.36	0.058
Blood tests	albumin (g/l)	39	34	<0.001	38	35	0.2	38	35	0.4	38	36	0.001
Blood tests	total bilirubin (µmol/L)	20	45	<0.001	21	26	0.3	22	9	<0.001	21	32	0.05
Blood tests	conjugated bilirubin (µmol/L)	12	33	<0.001	13	17	0.4	13	4	<0.001	13	23	0.029
Blood tests	unconjugated bilirubin (µmol/L)	10	15	<0.001	11	12	0.2	11	5	<0.001	11	10	>0.9
Blood tests	HCG (mIU/mL)	421.6	4,467.45	0.4	454.36	1.04	0.004	448.02	1	0.004	449.07	1.17	0.004
Blood tests	blood alcohol content g/l	0.83	0.48	<0.001	0.81	0.73	0.5	0.81	0.9	0.9	0.81	0.22	<0.001
Blood tests	lactate blood gas mg/dl	1.89	2.59	0.001	1.91	2.88	0.026	1.94	2.86	0.2	1.94	1.92	>0.9
Blood tests	LDH (IU/l)	342	382	0.2	344	378	0.3	344	557	0.6	346	327	0.6

Counts	number of symptoms	1.73	1.97	<0.001	1.74	1.76	0.8	1.74	1.98	0.1	1.74	1.98	<0.001
Counts	number of comorbidities	1.93	2.45	<0.001	1.96	1.77	0.028	1.96	2.39	0.035	1.95	2.46	<0.001
Counts	number of reasons for visiting ED	1.67	1.92	<0.001	1.68	1.85	0.009	1.68	1.83	0.3	1.68	1.83	0.006
Counts	number of habitual treatments	0.78	0.96	<0.001	0.79	0.61	<0.001	0.79	0.71	0.4	0.78	1.11	<0.001
Counts	number of prescriptions in ED	1.04	1.02	0.6	1.04	1	0.5	1.04	1.08	0.7	1.04	0.92	0.022
Temporal	tuesday admission	0.14	0.14	0.9	0.14	0.13	0.7	14%	14%	>0.9	0.14	0.14	>0.9
Temporal	monday admission	0.15	0.16	0.078	0.15	0.12	0.14	15%	14%	>0.9	0.15	0.17	0.4
Temporal	sunday admission	0.11	0.1	0.9	0.11	0.13	0.053	11%	12%	0.8	0.11	9.50%	0.5
Temporal	wednesday admission	0.14	0.13	0.4	0.14	0.15	0.7	14%	11%	0.5	0.14	0.13	0.8
Temporal	thursday admission	0.15	0.14	0.5	0.15	0.13	0.5	15%	14%	>0.9	0.15	0.15	>0.9
Temporal	friday admission	0.15	0.14	0.5	0.15	0.12	0.055	15%	16%	>0.9	0.15	0.14	>0.9
Temporal	saturday admission	0.12	0.12	>0.9	0.12	0.14	0.4	12%	7.90%	0.2	0.12	0.13	>0.9
Temporal	january admission	9.10%	9.60%	0.7	9.20%	8.00%	0.6	9.10%	10%	0.8	9.10%	9.70%	0.9
Temporal	december admission	7.60%	9.20%	0.028	7.70%	7.70%	>0.9	7.70%	9.40%	0.7	7.70%	8.00%	>0.9
Temporal	february admission	9.10%	0.11	0.003	9.20%	0.1	0.7	9.20%	13%	0.2	9.20%	0.11	0.3
Temporal	march admission	9.50%	0.1	0.8	9.60%	0.1	>0.9	9.60%	7.90%	0.7	9.60%	8.90%	0.8
Temporal	april admission	0.1	8.40%	0.035	0.1	7.00%	0.023	9.90%	8.90%	0.9	10.00%	9.10%	0.7
Temporal	may admission	9.10%	7.90%	0.14	9.00%	9.50%	>0.9	9.00%	7.30%	0.7	9.00%	8.40%	0.8
Temporal	june admission	6.90%	5.10%	0.005	6.90%	5.80%	0.5	NA	NA	NA	6.90%	5.60%	0.4
Temporal	july admission	6.90%	6.70%	>0.9	6.90%	5.30%	0.2	6.90%	5.20%	0.7	6.90%	7.50%	0.8
Temporal	august admission	7.20%	7.20%	>0.9	7.20%	7.50%	>0.9	7.20%	6.30%	0.9	7.20%	8.30%	0.5
Temporal	october admission	6.40%	6.00%	0.8	6.30%	8.00%	0.2	6.40%	4.70%	0.6	6.40%	6.10%	>0.9
Temporal	november admission	6.70%	6.20%	0.6	6.70%	7.40%	0.7	6.70%	9.40%	0.3	6.70%	6.00%	0.8
Temporal	morning admission	0.21	0.22	0.3	0.21	0.18	0.3	21%	21%	>0.9	0.21	0.22	0.6
Temporal	afternoon admission	0.39	0.41	0.5	0.4	0.34	0.005	39%	33%	0.2	0.39	0.44	0.02
Temporal	evening admission	0.27	0.24	0.004	0.27	0.29	0.5	27%	27%	>0.9	0.27	0.23	0.074
Temporal	night admission	7.60%	6.90%	0.5	7.50%	0.11	<0.001	7.50%	6.80%	>0.9	7.60%	4.90%	0.025
Temporal	hour admission	14	13.8	0.09	14	13.8	0.3	14	14.1	0.8	14	14.1	0.7
Temporal	age of patient	65	77	<0.001	66	60	<0.001	66	70	0.002	66	76	<0.001
Temporal	length of stay in ED (hours)	5.9	5.8	0.7	5.9	4.2	<0.001	5.9	4.9	0.023	5.9	7.2	0.031
Other	directed to vital emergency room	0.11	0.22	<0.001	0.11	0.37	<0.001	12%	25%	<0.001	0.12	0.12	>0.9

Supplementary table 2 : Variable weight in each model
The weight of each variable was computed by computing the mean decrease impurity. Values were normalized on a scale from 0 to 100, representing the minimum and the maximum importance in each model.

Variable category	Variable	Death	Intubation	CPR	Palliative care
Patient history	stroke	7	0	3	8
Patient history	hypertension	3	9	11	1
Patient history	heart attack	5	0	3	6
Patient history	arrhythmia	6	1	12	9
Patient history	COPD	5	11	0	3
Patient history	diabete	3	7	1	2
Patient history	addiction	6	7	0	1
Patient history	depression	3	7	5	2
Patient history	ulcer	6	3	0	6
Patient history	cancer	5	8	6	5
Usual treatments	antihypertensive	3	1	3	2
Usual treatments	diuretic	6	2	7	2
Usual treatments	antithrombotic	3	4	1	1
Usual treatments	anti-ulcer	5	8	1	4
Usual treatments	lipid lowering	3	2	16	4
Usual treatments	psychotropic	3	2	3	8
Usual treatments	painkillers	3	0	0	6
Reasons for admission	malaise	3	0	11	2
Reasons for admission	confusion	5	4	2	5
Reasons for admission	headache	2	7	1	2
Reasons for admission	dyspnea	13	1	2	4
Reasons for admission	cough	4	2	8	2
Reasons for admission	fever	3	0	1	2
Reasons for admission	weakness	5	1	2	4
Reasons for admission	chest pain	3	2	9	3
Reasons for admission	abdominal pain	1	4	0	4
Reasons for admission	fall	3	3	1	2
Reasons for admission	dyspnea	14	9	1	4
Vital signs	systolic blood pressure (mmHg)	81	43	67	43
Vital signs	diastolic blood pressure (mmHg)	73	62	39	45
Vital signs	mean blood pressure (mmHg)	63	45	98	40
Vital signs	heart rate (BPM)	73	35	48	37
Vital signs	température (°C)	65	49	69	37
Vital signs	oxygen saturation (%)	48	58	38	53
Vital signs	Glasgow coma scale	57	21	4	20

Vital signs	respiratory rate (breaths per minute)	88	100	39	44
Vital signs	pain scale	68	41	47	66
Vital signs	capillary hemoglobin(g/dl)	54	50	60	50
Vital signs	blood sugar (mg/dl)	73	54	41	61
Vital signs	ketonemia (mmol/l)	57	46	19	57
Symptoms reported by doctors	confusion	4	3	1	5
Symptoms reported by doctors	headache	2	2	1	1
Symptoms reported by doctors	skin wound	5	3	0	1
Symptoms reported by doctors	cough	4	9	8	2
Symptoms reported by doctors	fever	2	7	11	2
Symptoms reported by doctors	mottled skin	15	1	3	8
Symptoms reported by doctors	weakness	4	5	2	4
Symptoms reported by doctors	chest pain	3	4	5	1
Symptoms reported by doctors	abdominal pain	2	0	1	3
Symptoms reported by doctors	fall	3	6	6	3
Complementary tests	cerebral scan	4	4	14	5
Complementary tests	chest scan	4	10	27	6
Complementary tests	abdominal scan	2	4	0	2
Complementary tests	electrocardiogram	3	3	3	2
Complementary tests	chest x-ray	4	6	7	2
Complementary tests	pelvis x-ray	3	0	0	4
Complementary tests	COVID-19 PCR	3	1	0	4
Prescriptions	activated clotting time	3	17	4	4
Prescriptions	pain monitoring	2	9	2	1
Prescriptions	temperature monitoring	2	1	8	1
Prescriptions	IV line mentionned	5	28	20	4
Prescriptions	vital signs monitoring	2	6	3	1
Prescriptions	neurological status monitoring	2	2	11	1
Prescriptions	catheter	2	7	13	1
Prescriptions	urine strip	3	5	1	2
Prescriptions	administration of usual treatments	3	11	10	2
Prescriptions	IV rehydration	2	8	1	1
Prescriptions	aerosol	8	4	8	2
Prescriptions	oral hydration	3	1	4	2
Prescriptions	painkiller	2	0	1	2
Prescriptions	antihypertensive	6	1	1	3
Prescriptions	diuretic	9	1	1	1
Prescriptions	antithrombotic	5	5	0	1

Prescriptions	antidiabetic	3	7	15	3
Prescriptions	antibiotique	4	4	4	2
Prescriptions	o2 (l/min)	99	41	74	55
Blood tests	red blood cells count (million/mm3)	49	42	47	42
Blood tests	platelets count (cells/mm3)	47	58	22	55
Blood tests	white blood cells count (million/mm3)	48	38	86	45
Blood tests	neutrophils count (million/mm3)	63	49	53	64
Blood tests	eosinophils count (million/mm3)	33	47	30	30
Blood tests	basophils count (million/mm3)	27	69	73	22
Blood tests	mean corpuscular RBC volume (fl)	33	37	15	27
Blood tests	mean corpuscular platelets volume (fl)	26	30	20	25
Blood tests	mean corpuscular hemoglobin (g/dl)	27	44	42	23
Blood tests	mean corpuscular hemoglobin concentration (g/dL)	26	41	6	27
Blood tests	red cell distribution width (%)	37	39	48	32
Blood tests	quick test patient (s)	50	38	18	33
Blood tests	prothrombin (%)	58	49	36	31
Blood tests	international normalized ratio (no unit)	51	48	55	30
Blood tests	activated partial thromboplastin time (s)	36	29	58	33
Blood tests	APTT ratio (no unit)	43	35	71	39
Blood tests	D-dimer (mg/dl)	71	58	83	60
Blood tests	fibrinogen (g/l)	69	41	38	60
Blood tests	immature granulocytes (million/l)	78	70	36	64
Blood tests	factor V(%)	61	57	43	41
Blood tests	thrombin time patient (s)	100	10	25	53
Blood tests	Vitamin B9 (nmol/L)	68	50	25	41
Blood tests	factor II (%)	54	64	15	42
Blood tests	factor VII (%)	55	73	46	37
Blood tests	factor X (%)	55	49	43	43

Blood tests	natremia (mmol/l)	58	52	82	46
Blood tests	kalemia (mmol/l)	76	84	100	64
Blood tests	chloride (mmol/l)	52	51	43	40
Blood tests	bicarbonate (mmol/l)	46	65	59	37
Blood tests	glycemia (mmol/l)	47	69	67	33
Blood tests	urea (mmol/l)	70	66	43	51
Blood tests	uric acid (µmol/l)	85	41	70	81
Blood tests	protidemia (g/l)	61	57	33	48
Blood tests	calcium (mmol/l)	60	66	33	54
Blood tests	phosphorus (mmol/l)	82	70	40	49
Blood tests	CRP mg/l	51	67	34	71
Blood tests	BNP (pg/ml)	95	58	30	44
Blood tests	troponin (ng/l)	65	58	50	47
Blood tests	CKD clearance (ml/min)	78	45	30	49
Blood tests	creatine kinase (IU/l)	59	63	42	45
Blood tests	ASAT (IU/L)	51	21	98	28
Blood tests	ALAT (IU/L)	45	54	29	37
Blood tests	alkaline phosphatase (IU/L)	77	67	61	95
Blood tests	gamma gt (IU/L)	50	49	73	50
Blood tests	lipase (IU/L)	52	50	83	64
Blood tests	pH (no units)	60	48	21	42
Blood tests	pCO2 (mmHg)	44	47	24	62
Blood tests	total CO2 (mmol/L)	54	89	16	49
Blood tests	pO2 (mmHg)	54	58	48	37
Blood tests	oxygen saturation (%)	40	45	69	29
Blood tests	hemoglobin (g/dl)	43	34	38	45
Blood tests	oxygen carrying capacity (ml/dl)	45	90	68	40
Blood tests	base excess (mEq/l)	65	49	52	45
Blood tests	albumin (g/l)	93	49	75	63
Blood tests	total bilirubin (µmol/L)	54	36	51	48
Blood tests	conjugated bilirubin (µmol/L)	54	51	60	39
Blood tests	unconjugated bilirubin (µmol/L)	51	47	69	47
Blood tests	HCG (mIU/mL)	49	46	47	28
Blood tests	blood alcohol content g/l	63	38	29	47
Blood tests	lactate blood gas mg/dl	67	30	33	36
Blood tests	LDH (IU/l)	70	88	72	34
Counts	number of symptoms	21	26	42	25

Counts	number of comorbidities	22	17	13	43
Counts	number of reasons for visiting ED	24	27	27	17
Counts	number of habitual treatments	23	53	12	38
Counts	number of prescriptions in ED	24	56	28	18
Temporal	tuesday admission	5	5	0	3
Temporal	monday admission	5	2	3	6
Temporal	sunday admission	3	5	1	6
Temporal	wednesday admission	3	0	1	2
Temporal	thursday admission	5	9	1	5
Temporal	friday admission	4	3	6	3
Temporal	saturday admission	4	1	7	4
Temporal	january admission	3	9	1	7
Temporal	december admission	7	8	0	5
Temporal	february admission	9	0	3	7
Temporal	march admission	7	9	15	7
Temporal	april admission	3	11	3	8
Temporal	may admission	6	16	2	3
Temporal	june admission	1	1	NA	4
Temporal	july admission	3	0	0	3
Temporal	august admission	3	11	0	7
Temporal	october admission	6	4	14	4
Temporal	november admission	4	11	11	5
Temporal	morning admission	6	7	6	5
Temporal	afternoon admission	3	7	11	1
Temporal	evening admission	2	2	6	2
Temporal	night admission	5	14	4	3
Temporal	hour admission	48	47	24	40
Temporal	age of patient	66	26	39	50
Temporal	length of stay in ED (hours)	95	77	88	100
Other	directed to vital emergency room	18	39	41	4

8.2. Lien vers le code R de l'article 3

L'intégralité du script correspondant à l'article 3 a été partagé en ligne.

Le code R a été revu et modifié en regard des rapports avant soutenance. Il a été partagé en ligne sur la plateforme GitHub au format « R Makdown ». Le fichier du code est disponible en cliquant sur le lien suivant :
https://github.com/rafi-gangloff/PREDICT-IA/blob/85f6aea142646eb5a63cd303d6e71e98aabc75fd/PREDICT%20IA%20RAFI%20V2%202.Rmd
Le bouton de de téléchargement est celui-ci :  . L'ouverture du fichier sous R Studio permet de visualiser en partie gauche le script et en partie droite le plan après clic sur le bouton « outline ». Les étapes principales de l'étude, incluant celles de la modélisation, y apparaissent désormais clairement pour le lecteur. Les commentaires inutiles ont été supprimés.

Concernant la variable « nb jour » :

- L'extraction de plusieurs années de compte rendus pour un CHU correspond à une masse trop importante de données pour être collectées en une seule fois. Il a donc été décidé de collecter les données par blocs de 1 jour au moyen d'une boucle répétée selon la durée en jours de l'étude. « nb jours » est la variable « durée en jours de l'étude ». Elle est calculée en soustrayant la date de début à la date de fin de l'étude et en y ajoutant 1 jour.

- Exemple pour une étude durant 2 jours : Si DATE DE DEBUT= 01/05/2023 et DATE DE FIN = 02/05/2023, DATE DE DEBUT-DATE DE FIN renvoie une valeur de 1 jour, d'où l'implémentation de +1 qui vise à corriger ce problème.

Titre : Prédiction par apprentissage automatique des évènements critiques présentés par les patients admis aux urgences

Mots-clés : apprentissage automatique, urgences, soins critiques, données massives

Résumé : La mise en adéquation des moyens déployés avec le niveau de gravité de chaque patient est un enjeu important pour la pérennisation et l'amélioration du système de santé. L'essor simultané de l'interopérabilité des bases de données permettant la collecte de données normalisées et des statistiques prédictives ouvre de nouvelles perspectives dans ce domaine en laissant supposer qu'il serait possible d'entraîner des modèles prédictifs permettant à la fois d'améliorer le service rendu aux patients en individualisant leur parcours de soin tout en rationalisant l'effort collectif nécessaire à leur prise en charge. Cette thèse avait pour objectif de prédire la survenue d'évènements critiques chez les patients ayant recours aux urgences en utilisant des modèles basés sur l'apprentissage automatique. Trois articles composent ce travail de recherche. Le premier visait à détecter les arrêts cardiaques préhospitaliers en utilisant les enregistrements vocaux des appels des témoins au SAMU. Dans cette optique, plusieurs modèles ont été développés pour détecter les arrêts cardiaques en se basant sur les caractéristiques acoustiques de la voix de l'appelant. Le deuxième article se concentrait sur l'optimisation du diagnostic de COVID-19 en intégrant les tests diagnostiques de référence de type RT-PCR à d'autres éléments clinico biologiques. Les modèles d'apprentissage automatique développés permettaient une augmentation des performances diagnostiques en ce contexte de pandémie débutante dans l'hypothèse d'une stratégie « zéro-COVID ». Le troisième article avait pour objectif la prédiction de quatre évènements critiques chez les patients hospitalisés après un passage aux urgences : la survenue d'un décès, la nécessité d'intubation, la réanimation cardiopulmonaire et la décision de réaliser des soins palliatifs. Des modèles de forêts aléatoires y ont été développés en intégrant des données les plus exhaustives possible afin d'établir un profil détaillé des patients : dates d'admission, temps de passage, antécédents, observations médicales, constantes vitales, examens biologiques et comptes-rendus d'imagerie. Une excellente performance pour la prédiction des quatre événements d'intérêt a été retrouvée dans cet article. Des limites ont par ailleurs été identifiées, comme la nécessité de valider ces approches dans des contextes cliniques réels et d'explorer davantage leur interprétabilité en pratique quotidienne. Ce travail de recherche apporte une contribution significative pour la prédiction des évènements critiques et permettra le développement d'applications visant à améliorer le parcours de soin des patients confrontés à une situation clinique susceptible de mettre en jeu leur pronostic vital.

Title: Machine learning modeling of critical events in emergency situations: from pre-hospital prediction to hospital care pathway optimization

Keywords: Machine learning, Critical care, Emergency Medicine, Big Data

Abstract. The challenge of adjusting the resources deployed to match the level of severity for each patient is an important issue for the sustainability and improvement of the healthcare system. The simultaneous development of database interoperability, enabling the collection of standardized data, and of predictive statistics is opening up new prospects in this field, suggesting that it may be possible to train predictive models to improve the service provided to patients by individualizing their journey through the care system, while rationalizing the collective effort required for their care. The aim of this thesis was to predict the occurrence of critical events in emergency patients, using models based on machine learning. Three articles are included in this work. The first was aimed at detecting pre-hospital cardiac arrest using voice recordings of witness calls to the SAMU. To this end, several models were developed to detect cardiac arrest based on the acoustic characteristics of the caller's voice. The second article focused on optimizing the diagnosis of COVID-19 by integrating reference RT-PCR diagnostic tests with other clinico-biological elements. Machine learning models were developed to improve diagnostic performance in the context of an emerging pandemic, based on the hypothesis of a "zero-COVID" strategy. The objective of the third article was to predict four critical events in hospitalized patients after a passage through the emergency department: death, intubation, cardiopulmonary resuscitation and palliative care. Random forest models were developed, integrating the most exhaustive data possible to provide in-depth patient profiling: admission dates, length of stay, history, medical observations, vital signs, biological examinations and imaging reports. This article demonstrated an excellent performance in predicting the four events of interest. Limitations were also identified, such as the need to validate these approaches in real clinical settings. This research work makes a significant contribution to the prediction of critical events and will enable the development of applications designed to improve the care of patients faced with life-threatening clinical situations.